Diffusion Mental Averages

Phonphrm Thawatdamrongkit Sukit Seripanitkarn Supasorn Suwajanakorn
VISTEC, Thailand

Figure 1. Our method unveils the *Mental Averages* encoded by pre-trained diffusion models across diverse concepts (left) and generalizes across different model variants (right), offering new tools for analyzing model biases and interpreting learned concept representations.

Abstract

Can a diffusion model produce its own "mental average" of a concept—one that is as sharp and realistic as a typical sample? We introduce Diffusion Mental Averages (DMA), a model-centric answer to this question. While prior methods aim to average image collections, they produce blurry results when applied to diffusion samples from the same prompt. These data-centric techniques operate outside the model, ignoring the generative process. In contrast, DMA averages within the diffusion model's semantic space, as discovered by recent studies. Since this space evolves across timesteps and lacks a direct decoder, we cast averaging as trajectory alignment: optimize multiple noise latents so their denoising trajectories progressively converge toward shared coarse-to-fine semantics, yielding a single sharp prototype. We extend our approach to multimodal concepts (e.g., dogs with many breeds) by clustering samples in semantically-rich spaces such as CLIP and applying Textual Inversion or LoRA to bridge CLIP clusters into diffusion space. This is, to our knowledge, the first approach that delivers consistent, realistic averages, even for abstract concepts, serving as a concrete visual summary and a lens into model biases and concept representation. Please visit our project page: `https://diffusion-mental-averages.github.io`

1. Introduction

When most people imagine a bird, they picture something small and ordinary, perhaps a sparrow rather than a large ostrich or the exotic Nene. That mental prototype reflects our exposure, preferences, and cultural context. Diffusion models can generate countless variations of a single concept, yet one might ask whether they, too, hold a single "mental average" that defines what they consider typical. Visualizing such prototypes provides a tangible summary of a prompt, reveals biases in the model and its training data, and offers new ways to analyze or steer its generative behavior.

To create a concept prototype learned by diffusion models like StableDiffusion [76], we can begin by generating many samples of the same concept. Prior work visually summarizes such collections by selecting representative samples according to some criteria [15, 40, 85], or by identifying discriminative patches [20, 36, 46, 84]. Data Distillation [12, 13, 89] instead synthesizes a compact set of prototypes by distilling information from the entire image collection. Our goal, however, is to synthesize a single prototype that reconciles cross-sample variations, which selection-based methods cannot achieve. Unlike Data Distillation, which targets downstream objectives and often yields unnatural or inconsistent prototypes, we target concept summarization via prototypes that match the sharpness and realism of typical samples from the probe model.

A common way to synthesize such prototypes is by averaging pixels across *spatially aligned* samples [26, 39, 56, 70, 106]. However, spatial alignment is rarely feasible beyond constrained domains like faces [18, 70, 90] or animals [70], and virtually impossible for abstract concepts like "poverty." Even with perfect alignment, fine details tend to average out at the pixel level, yielding unrealistic appearance [56, 70]. A promising direction is to perform averaging in a semantic space, where distances reflect conceptual similarity [37]. This can be done by encoding images into the semantic space, averaging their embeddings, then decoding—a straightforward process in autoencoders [2, 3], VAEs [43], or GANs [28, 37] equipped with inversion mechanisms [107]. Diffusion models, however, lack such an explicit bottleneck representation. Although some studies have identified "semantic layers" within diffusion models [44], the semantic information is distributed across timesteps, with no direct way to decode a time-averaged embedding into a realistic average.

To address this, we recast averaging as trajectory alignment within the diffusion model. Instead of generating samples then averaging them post hoc, we jointly optimize multiple noise latents so that their denoising trajectories progressively converge toward shared coarse-to-fine semantics. Conceptually, the process first aligns high-level structures, such as global shape and layout, toward a semantic mean, then gradually refines alignment to local patterns, enabling spatial averaging of fine details without sacrificing realism. During optimization, each sample follows its own denoising trajectory while being constrained to match the mean semantic representation of all samples at each timestep. The resulting trajectory forms a semantic consensus across samples, which can then be decoded into a single sharp and realistic prototype—effectively the model's own "mental average" of the concept.

While averaging captures a concept's shared structure, many real-world concepts are visually multimodal or ambiguous: "dog" spans many breeds, while "bike" may mean a bicycle or a motorbike. In such cases, a single average may blur distinct modes and obscure diversity. To capture this, we create mode-specific prototypes by clustering samples in a semantic space and averaging within each cluster. We find that CLIP [74] serves as a stable and semantically rich space, leading to clearer mode separation than diffusion-based feature spaces. Clustering can also be grounded using models like BLIP [50], enabling user-defined partitioning (e.g., by ethnicity for human faces). However, the latent space used for clustering (e.g., CLIP) differs from the one used for averaging within the diffusion model, causing inconsistencies between clusters and their prototypes. To bridge this gap, we adapt the conditioning of the diffusion model to be cluster-specific: for each cluster, we learn a textual inversion embedding [24] or LoRA [34] that captures its semantic subspace and guides the denoising process toward that region, yielding more faithful averages.

We evaluate our method for generating concept averages across diverse categories, including animals, humans, objects, and abstract concepts, and compare it against several baselines, including GANgealing [70] and adapted dataset distillation methods such as D^4M [89] and MGD^3 [13]. Our method produces results that are both visually realistic and consistent across different initial random seeds, whereas the baselines tend to be either consistent but blurry or realistic but inconsistent. We further demonstrate that our mode discovery method can effectively generate averages for multiple modes in both unsupervised and grounded setups, and that our approach generalizes to other diffusion architectures, such as DiT [69]. In summary, our contributions are:

- We propose the first method to produce consistent and realistic averages directly from pre-trained diffusion models by formulating the task as trajectory alignment.
- We demonstrate that DMA generalizes to multimodal concepts with light-weight model conditioning, as well as across diffusion model variants and architectures.

2. Related Work

Prototype image. A prototype image aims to represent the visual essence of a concept or dataset. Early approaches decompose images into patches, group them by visual similarity, and use these clusters as prototypical elements [19, 20, 27, 36, 46, 51, 75, 84, 86]. While effective for analyzing local features, patch-based prototypes fail to capture holistic relationships between visual components. Other works instead select entire images to serve as prototypes [10, 15, 40, 85]. However, this approach is limited by the finite sample set, potentially overlooking variations or information contained in other samples.

Beyond selecting existing samples, another line of work focuses on synthesizing prototype images. Dataset distillation methods [11, 12, 16, 47, 93, 98, 103–105] generate synthetic images that encapsulate the knowledge of an entire class, but their prototypes often lack recognizable structure and visual coherence. Recent methods in dataset distillation employ diffusion models to distill class-level representations into a set of high-quality prototypes [13, 29, 89]. However, they primarily target downstream tasks over conceptual summarization and can exhibit inconsistent generations. Pixel-space averaging on spatially aligned images provides an intuitive way to summarize visual information [26, 56, 106]. Although alignment methods [5, 45, 62, 65, 70] preserve the overall structure of the average image, fine details are averaged out and blurred because pixel-space averaging operates on low-level image features, such as texture and illumination, which vary significantly across samples. A more effective strategy is to perform averaging in a semantic latent space, where high-level concept information is encoded. This can be done by encoding samples into the latent space, computing their mean representation, and decoding it back into the image domain [2, 3, 28, 37, 43]. However, this remains unexplored in the context of diffusion models, which lack an explicit semantic latent space and a dedicated decoder for reconstruction.

Building on this idea, we aim to construct high-quality average images of any concept by leveraging the semantic space of pre-trained diffusion models, without requiring access to the underlying dataset. Closest to our work, Feng et al. [22] use compositional generation [55] in a diffusion model fine-tuned with GPS signals to produce average representations for specific geographic areas, relying on multiple location-based conditions. In contrast, we operate in text-to-image diffusion models, where such multi-conditional inputs (e.g., GPS coordinates) are impractical, as each concept is defined solely by a single text prompt.

Semantic space in diffusion models. Our method leverages the semantic spaces of UNet–based diffusion models to generate average images. Some studies have explored these spaces for tasks such as segmentation [4, 52, 59, 60, 64, 71, 78, 92, 96], visual correspondence [57, 60, 91, 99], and concept ablation [7, 8]. However, these semantic spaces typically retain spatial information, suitable for their tasks but suboptimal for ours, as they can introduce the same blurring issues seen in pixel-space averaging. Other works attempt to create high-level semantic spaces, as in autoencoders or VAEs [43], by using an encoder to compress images into semantic latent vectors that guide the denoising process [35, 48, 72, 94]. However, this requires retraining the entire network and is unsuitable for our goal of finding the mental average within a pre-trained diffusion model. Kwon et al. [44] show that the U-Net bottleneck layer (h-space) exhibits properties, such as linearity, that make it an effective semantic representation for our task. Semantic information in diffusion models is also found to spread across timesteps, with early steps capturing coarse structure and later steps encoding fine details [14, 17, 68, 80]. We leverage the h-space for its low spatial specificity and rich semantic content.

Noise optimization in diffusion models. Once the average semantic latent is obtained, decoding it into an average image is non-trivial. A naive approach is to directly substitute the h-space latent with the average during denoising, but this often produces inconsistent outputs because skip connections allow unconstrained information from other layers to influence the output. To constrain activations in other layers without modifying model parameters, noise optimization has proven effective. Prior works have employed this approach for enhancing image quality [21, 30, 41, 73], generating rare concepts [79], and motion editing [38, 54, 66].

We adopt noise optimization to decode the average semantic latent into an average image by applying it sequentially across diffusion timesteps.

3. Background on T2I Diffusion Models

Modern text-to-image (T2I) diffusion models [33, 87] can be viewed as discrete-time instances within the broader family of score-based [88] and flow-matching [53] generative models, which learn to transform Gaussian noise into data through a reverse stochastic or deterministic process. These models generate high-resolution images efficiently by operating in a latent space rather than directly in pixel space [76]. A Variational Autoencoder (VAE) [43] first maps images into a compact latent space, where the diffusion process is applied.

For text-conditioned generation, a prompt c is encoded by a transformer-based text encoder to guide the denoising trajectory. At inference, generation begins from a Gaussian latent $\mathbf{z}_T \sim \mathcal{N}(0, I)$ and progressively denoises it into a clean latent $\mathbf{z}_0$, which is decoded by the VAE decoder into the final image. At each timestep t, the model predicts the noise $\epsilon_t = \epsilon_\theta(\mathbf{z}_t, t, c)$ and the next latent $\mathbf{z}_{t-1}$ is computed using, e.g., the DDIM update [87]:

$$\mathbf{z}_{t-1} = \sqrt{\alpha_{t-1}} \left(\frac{\mathbf{z}_t - \sqrt{1-\alpha_t}\,\epsilon_t}{\sqrt{\alpha_t}} \right) + \sqrt{1-\alpha_{t-1}}\,\epsilon_t. \quad (1)$$

To improve image-prompt fidelity, Ho and Salimans [32] introduced classifier-free guidance (CFG), which blends conditional and unconditional predictions:

$$\tilde{\epsilon}_\theta(\mathbf{z}_t, t, c) = (1-w)\,\epsilon_\theta(\mathbf{z}_t, t, c) - w\,\epsilon_\theta(\mathbf{z}_t, t), \quad (2)$$

where w is the guidance scale controlling the trade-off between text faithfulness and sample diversity.

4. Method

Given a concept prompt, we aim to synthesize an average image that captures the concept's shared semantics under a probe diffusion model while preserving visual realism.

Averaging across diffusion samples is conceptually simple yet technically challenging: pixel-space averaging destroys realism, and feature-space averaging is ill-defined because diffusion models lack a semantic bottleneck and decoder. Semantic information is instead distributed along the

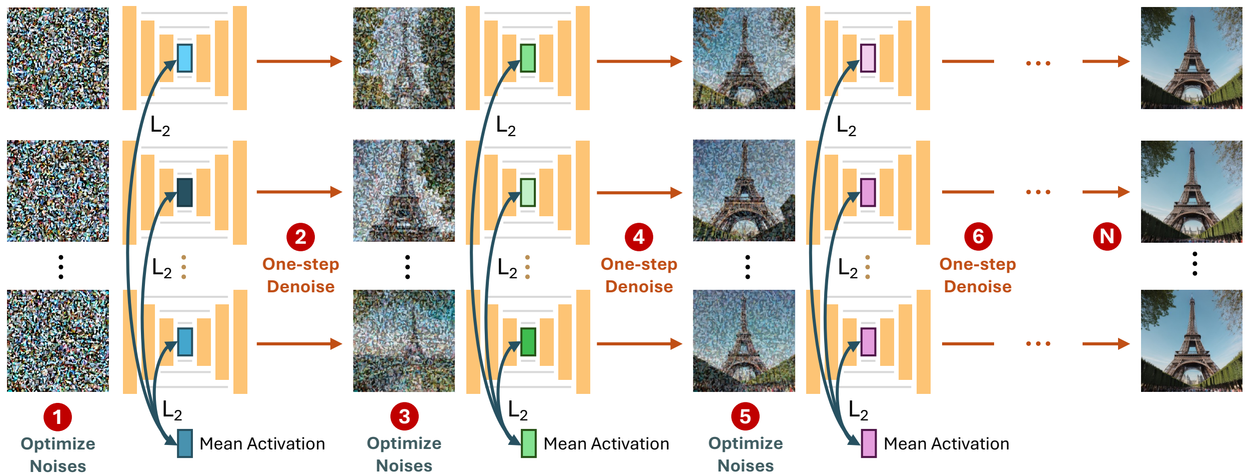

Figure 2. **Overview.** Multiple noise latents are jointly optimized so that their denoising trajectories converge toward shared semantics. At each timestep t, their h-space activations are averaged to form a semantic target, and each latent is optimized to match it before denoising to the next step. Repeating this process across timesteps aligns coarse-to-fine semantics, yielding a single "mental average" of the concept.

denoising trajectory, evolving from coarse layout to fine detail. This motivates performing averaging not in image or latent space confined to a single layer, but along the model's denoising process itself. We therefore recast the problem as aligning multiple denoising trajectories so that they converge toward a shared semantic consensus. Figure 2 provides an overview.

To achieve this, we leverage a semantically meaningful latent layer, referred to as the h-space, across *multiple* diffusion timesteps. This layer, located near the middle of the U-Net denoiser, has been shown to encode interpretable and approximately linear semantics [44, 49, 67, 83]. Rather than collapsing h-space representations from all timesteps into a single latent vector, which would be non-decodable, we progressively align the h-space representations of multiple noise latents to their means along the denoising process. This mirrors the diffusion model's inherent generation from high-level to low-level attributes, enabling the model to first reach agreement on global structure, such as pose and composition, before successively aligning lower-level attributes.

Specifically, we start with K noisy latents $\{\mathbf{z}_k^{(0)}\}_{k=1}^K$, each drawn from $\mathcal{N}(0, \mathbf{I})$. At each diffusion timestep t, we perform the following steps:

1. Compute the average h-space activation. For each noisy latent $\mathbf{z}_k^{(t)}$, we feed it into the diffusion model conditioned on timestep t and the prompt to extract its h-space activation $H(\mathbf{z}_k^{(t)})$, and compute the average activation across all K latents, denoted as $\bar{A}^{(t)} = \frac{1}{K}\sum_{k=1}^K H(\mathbf{z}_k^{(t)})$.

2. Noise optimization. To align each noisy latent with the average activation $\bar{A}^{(t)}$, we optimize each $\mathbf{z}_i^{(t)}$ using the following objective:

$$\min_{\mathbf{z}_i^{(t)}} \left\| H(\mathbf{z}_i^{(t)}) - \bar{A}^{(t)} \right\|_2^2 . \tag{3}$$

Minimizing this loss gradually shifts each latent toward the shared semantic feature at timestep t. The optimization is performed for N iterations for each latent.

3. Denoising. After aligning the noisy latents with the mean semantic features, each optimized latent is denoised to the next timestep using DDIM sampling [87].

This process is repeated across diffusion timesteps until a cutoff $t_{\text{stop}} < T$, which helps reduce computation cost. Then, we continue denoising with standard DDIM sampling until completion and decode any final latent $\mathbf{z}_k^{(T)}$ with the VAE decoder to produce the prototype image, as outlined in Algorithm 1. We examine how different noise latents converge under varying cutoff values t_{stop} in Appendix C.4.

4.1. Mode Discovery

While our method aligns multiple trajectories toward a single semantic consensus, this process may oversimplify concepts that the diffusion model represents through several distinct semantic modes. For example, "crane" can refer to a bird or a construction machine—two modes that are difficult to meaningfully blend into a single prototype.

We extend our method with a simple mode separation step: we first cluster the noise latents and then compute an average image per cluster. Effective mode discovery requires a feature space that captures high-level semantics. Although the internal h-space encodes semantic content, its representations are stochastic and vary across timesteps, making it unreliable for consistent mode separation. Instead, we use the CLIP embedding space [74], which offers a stable, high-level semantic representation [6, 9, 25, 77, 82]. To support user-specified clustering (e.g., by ethnicity for "doctor"), we alternatively employ a grounded approach using BLIP-VQA [50] to obtain attribute-focused embeddings [56] for targeted mode discovery. Noise latents are

Algorithm 1 Diffusion Mental Averaging

Require: Diffusion model $\mathcal{D}$, concept prompt c, optimization cutoff t_{stop}, diffusion timestep schedule $\{\tau_t\}_{t=1}^{T}$, number of sampled latents K, number of optimization steps N, learning rate η, h-space activation $H(\cdot)$

Ensure: Prototype Image I

Initialize a set of K noisy latents $\mathbf{Z}^{(0)} = \{\mathbf{z}_k^{(0)}\}_{k=1}^{K}$ with $\mathbf{z}_k^{(0)} \sim \mathcal{N}(0, \mathbf{I})$
for $t = 0$ **to** $T-1$ **do**
  if $t \leq t_{\text{stop}}$ **then**
    $\bar{A}_t \leftarrow \frac{1}{K} \sum_{\mathbf{z} \in \mathbf{Z}^{(t)}} H\big(\mathbf{z}; \mathcal{D}, \tau_t, c\big)$
    for $k = 1$ **to** K **do**
      for $i = 1$ **to** N **do**
        $\mathcal{L} \leftarrow \left\| H\left(\mathbf{z}_k^{(t)}; \mathcal{D}, \tau_t, c\right) - \bar{A}_t \right\|_2^2$
        Update $\mathbf{z}_k^{(t)}$ using Adam$(\nabla_{\mathbf{z}_k^{(t)}} \mathcal{L}, \eta)$
      end for
    end for
  end if
  $\mathbf{Z}^{(t+1)} \leftarrow$ DDIM-Sampling$\big(\mathbf{Z}^{(t)}; \mathcal{D}, \tau_t, c\big)$
end for
$I \leftarrow$ VAE-Decode$(\mathbf{z}_k^{(T)})$ for any k
Return: Prototype Image I

clustered by generating samples through standard denoising, embedding them with CLIP or BLIP-VQA, and applying a clustering method such as GMM; cluster assignments are then mapped back to the latents.

After clustering, a straightforward approach is to apply DMA independently within each cluster. However, we find that this often yields prototypes that do not faithfully reflect their cluster's semantics. We hypothesize that this inconsistency stems from a mismatch between the semantic space used for clustering and the diffusion model's h-space. Since these spaces encode semantics differently, samples that form a coherent mode in one space may scatter across the other, causing the average to lose semantic coherence.

To mitigate this, we adjust the model's conditioning to steer denoising toward each cluster's semantic subregion. This can be done using techniques such as Textual Inversion [24], which learns a compact textual embedding from a set of images, or by training a LoRA [34] to fine-tune the model's conditioning pathway. While Textual Inversion can handle simple or well-defined modes, we find that its limited capacity struggles with complex modes [102]. LoRA, being more expressive, captures richer intra-mode variation and provides better alignment between semantic modes and their corresponding prototypes. We compare the two approaches in our experiments. With these lightweight conditioning methods, our approach is able to summarize each concept's mode with greater semantic fidelity.

5. Experiments

Implementation details. Unless stated otherwise, we use a variant of Stable Diffusion called Realistic Vision v5.1 [81] due to its higher image quality. All experiments use classifier-free guidance (CFG) scale of 7.0 and 20 DDIM sampling steps. We set the number of latents $K = 1000$. At each diffusion timestep, each latent is optimized using Adam [42] with a learning rate of 2×10^{-2} for 300 iterations. We set the cutoff $t_{\text{stop}} = 10$. The optimization process took about 10 hours on an NVIDIA RTX 4080. Additional details are provided in Appendix A.

5.1. Baselines

As no method currently exists for averaging diffusion models, we compare DMA to adapted versions of two related approaches: congealing and dataset distillation. All methods use the same 1,000 noise latents or their samples.

Average VAE. Starting from noisy latents, we denoise them, compute their mean in VAE latent space, and decode it once into pixel space.

GANgealing [70]. We generate samples from the same noise latents and use GANgealing's pretrained Spatial Transformer Network to align the images. The aligned images are then averaged in pixel space.

D^{4}M [89] uses cluster centers in VAE space as prototypes and refines them by adding noise and denoising through a diffusion model. In our adaptation, we set IPC = 1, so the prototype becomes the average of all clean VAE latents. We inject noise into this averaged latent at timestep $t = 6$ (equivalent to the default SDEdit strength 0.7) and denoise once to obtain the prototype image.

MGD^3 [13] also uses cluster centers as prototypes but introduces Mode Guidance to steer predicted noise toward the prototype. In our adaptation, we also set IPC = 1, and apply Mode Guidance for the first 10 steps (default 50% of total steps) with the default mode guidance of 0.1, followed by standard denoising to obtain the prototype image. See our Appendix D for results using other hyperparameter values.

5.2. Prototypical Image Evaluation

We evaluate on 12 concepts spanning four categories (Animal, Person, Object, Abstract). For each concept, we generate 10 sets of 1,000 samples, and each method produces 10 prototypes per set, yielding 100 prototypes per concept per method. This evaluation does not perform clustering as in the multimodal setting (Sec. 4.1). We use three metrics:

1. Consistency (↓) measures how reliably a method converges to the same prototype when starting from different initializations. For each set, we compute the aver-

Figure 3. Qualitative comparison across methods: **GANgealing** [70], **Avg. VAE**, **D^{4}M** [89], **MGD^3** [13], and **DMA (Ours)**. Rows show concepts: *astronaut*, *dog*, *bicycle*, *freedom*. GANgealing cannot process abstract concepts, so its *freedom* cell is intentionally left blank.

age pairwise distance among the 10 generated prototypes using CLIP [74] cosine distance, DreamSim [23], and LPIPS [101]. We report consistency scores averaged over all sets in each category, where lower is better.

2. Representativeness (↓) assesses how well each prototype captures the semantic center of its concept distribution. For each prototype, the Representativeness score is the average distance to the 1,000 samples in its set. We use the same three distance functions as in the Consistency metric and report scores averaged across all sets in each category, where lower is better.

3. ImageReward [97] (↑) is a learned reward model trained on large-scale human preference data. This metric reflects human preference by jointly evaluating image quality and text–image alignment. For each concept set, we compute the average score across the 10 generated prototypes, and then report the average ImageReward over all sets within each category, where higher is better.

Quantitative scores are shown in Table 1, with qualitative results in Figure 3. Avg. VAE and GANgealing achieve perfect consistency scores of zero by design, as they always produce identical outputs given the same image set. However, their averages are blurry and lack realism. DMA is significantly more consistent than D^{4}M and MGD^3 with better consistency scores across all three distance metrics. DMA also achieves the best ImageReward and representativeness scores across all metrics, demonstrating more faithful and

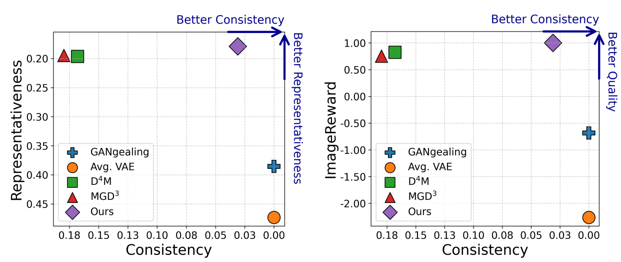

Figure 4. **Quality Trade-off.** DMA achieves a superior balance of representativeness, consistency, and perceptual quality.

Table 1. **Average scores** across four categories (*Animal*, *Person*, *Object*, *Abstract*) on Consistency (↓), Representativeness (↓), and ImageReward (↑). GANgealing† is evaluated only on three non-abstract categories, as it lacks pretrained GANs on abstract concepts

	Consistency (↓)			Representativeness (↓)			ImageReward (↑)
Method	CLIP	DreamSim	LPIPS	CLIP	DreamSim	LPIPS	
GANgealing† [70]	0	0	0	0.386	0.477	0.851	−0.684
Avg VAE	0	0	0	0.473	0.806	0.805	−2.262
D^4M [89]	0.168	0.274	0.572	0.197	0.363	0.672	0.823
MGD^3 [13]	0.180	0.319	0.643	0.195	0.364	0.687	0.755
DMA (Ours)	**0.031**	**0.032**	**0.129**	**0.179**	**0.341**	**0.655**	**1.002**

visually preferred concept summaries than all baselines.

Qualitatively, DMA generates clear, structured averages, for example, a detailed astronaut in uniform with averaged attributes, such as centered composition. By contrast, the baselines only capture vague helmet shapes and omit key components like the helmet's visor or uniform altogether. For abstract concepts, like "freedom," DMA consistently depicts the Statue of Liberty, whereas D^4M and MGD^3 generate inconsistent or distorted versions of the statue. GANgealing, meanwhile, cannot handle such concepts because no pretrained GAN exists for them. Furthermore, as shown in the trade-off plots in Figure 4, DMA delivers the most balanced performance, combining high consistency, strong representativeness, and perceptual realism.

5.3. Mode Averages

In this section, we present results of applying DMA to clustered images. To refine the model's conditioning for each mode, we use a CFG scale of 3.0 during inference and train a rank-1 LoRA for 2,000 steps with a learning rate of 10^{-4}. For Textual Inversion, we use a CFG scale of 7.0 during inference and train for 3,000 steps with a learning rate of 10^{-2}. For mode separation, we first reduce the computed features of CLIP-ViT-B/32 or BLIP-VQA-base [56] using PCA (to 2 dimensions for unsupervised clustering and 10 for grounded clustering), then cluster them using a Gaussian Mixture Model to obtain mode assignments.

Unsupervised clustering. Figure 5 presents our cluster averages discovered through unsupervised clustering, along with a single overall average, for four concepts. We observe that the single average is biased toward the dominant mode, thereby suppressing less frequent ones (e.g., the *fan* average only depicts an electric fan). In contrast, our cluster averages successfully reveal the distinct semantic modes within each concept.

Grounded clustering. Figure 6 presents our cluster averages from grounded clustering using BLIP-VQA. The results show that DMA can generate accurate average images across different grounded criteria. For the *car* concept

Figure 5. **DMA prototypes of unsupervised modes.** Top row: overall average. Bottom rows: averages of discovered modes.

grounded by color, the averages show that yellow and silver cars look similar, while red and blue cars form another consistent group. This demonstrates how DMA can be used as a probing tool to reveal how models organize and associate attributes within a concept.

LoRA and Textual Inversion. Figure 7 shows that using LoRA or Textual Inversion helps guide the model toward cluster-specific regions, producing averages that better reflect the modes' semantics, with LoRA better preserving color and car shape than Textual Inversion. This suggests that both methods help capture semantics beyond the h-space, including aspects typically encoded in the noisy latent, such as color and image layout [63, 95, 100].

5.4. Generalization Across SD Variants

We evaluate DMA across several Stable Diffusion variants: SD1.5, Realistic Vision v5.1, Dreamshaper PixelArt, and Flat-2D Animerge. In Figure 8, DMA can produce averages that are specific to each variant, revealing distinct stylistic or demographic interpretations. The concept *soldier* exhibits strong identity bias: SD1.5 and Realistic Vision consistently generate male soldiers, whereas PixelArt appears more gender-neutral, and Flat-2D Animerge ren-

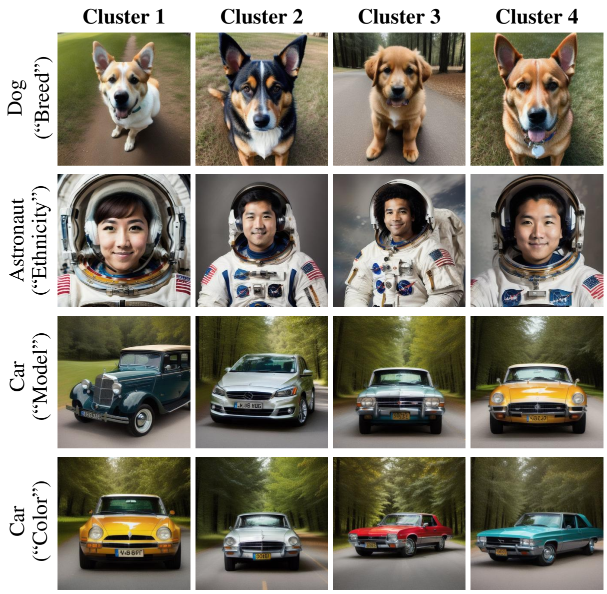

Figure 6. **DMA prototypes of grounded modes.** We cluster *astronaut* by ethnicity, *dog* by breed, and *car* by model or color.

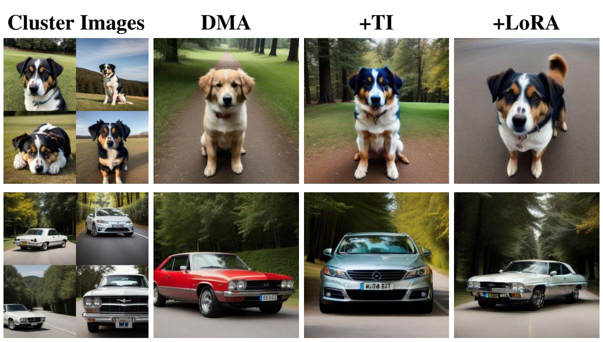

Figure 7. **LoRA & Textual Inversion** bridge the gap between the clustering and the h-spaces, enabling cluster-specific prototypes.

ders a female anime-style cadet. For *Italy*, all models converge to a similar "Venice canal" scene despite stylistic differences, indicating a shared dominant bias. This highlights that DMA generalizes across models and helps diagnose how fine-tuning alters or preserves concept semantics.

5.5. Generalization Across Architectures

We extend DMA to DiT-XL [69], a class-conditional, transformer-based diffusion model. Since DiT lacks a U-Net-style semantic bottleneck like the h-space, we instead use the output of the final transformer block, which yields the most consistent results (Figure 9). See Appendix A.4 and C.7 for implementation details and block comparisons.

6. Limitations and Discussion

As with sample generation, the concept and hyperparameters, like CFG, affect the quality and consistency of the average image. High-variation concepts require more samples for stable averages, while higher CFG improves prompt alignment but reduces diversity, influencing the final result.

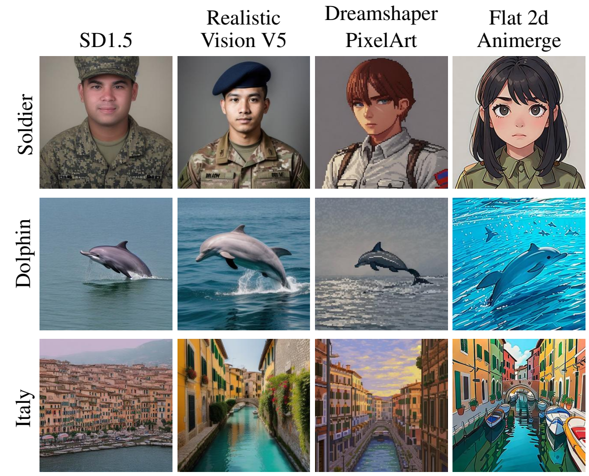

Figure 8. **DMA prototypes of various SD Variants.** *Soldier* inherits model-specific gender bias, while *Italy* exhibits a persistent static bias toward a Venice-like scene across all variants.

Figure 9. **DMA generalizes to the DiT architecture [69].**

Our clustering relies on external encoders and thus inherits their biases [1, 31, 58]; however, the averaging process remains unbiased once clusters are defined. Evaluating the representativeness score remains challenging as it depends on the specific embedding space, which can be subjective. We discuss this further in Appendix B.4. Extending DMA to other diffusion architectures requires identifying a semantic layer analogous to the h-space, which remains an open direction for future work. Currently, our method requires high computational costs, which are detailed in Appendix B.2. Further discussions regarding potential applications and limitations of our mode discovery are provided in Appendix B.1 and B.3.

Conclusion. We present a simple yet effective method for generating high-quality average images from pretrained diffusion models. By progressively aligning denoising trajectories across multiple noise latents, DMA produces consistent, representative results and extends to concepts with multiple meanings and modes. DMA outperforms baselines in consistency, image quality, and representativeness, and generalizes across diffusion models and architectures.

Acknowledgment. This research was supported by Thailand Science Research and Innovation (TSRI) Fundamental Fund (Grant number: FRB690039/0457; Project number: 4823418), SCB Public Company Limited, and PTT Public Company Limited.

References

[1] Vatsal Baherwani and Joseph James Vincent. Racial and gender stereotypes encoded into clip representations. In *The Second Tiny Papers Track at ICLR 2024*, 2024. 8

[2] Pierre Baldi. Autoencoders, unsupervised learning, and deep architectures. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 37–49. JMLR Workshop and Conference Proceedings, 2012. 2, 3

[3] Dor Bank, Noam Koenigstein, and Raja Giryes. Autoencoders. *Machine learning for data science handbook: data mining and knowledge discovery handbook*, pages 353–374, 2023. 2, 3

[4] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021. 3

[5] Nir Barel, Ron Shapira Weber, Nir Mualem, Shahaf E Finder, and Oren Freifeld. Spacejam: a lightweight and regularization-free method for fast joint alignment of images. In *European Conference on Computer Vision*, pages 180–197. Springer, 2024. 3

[6] Manuele Barraco, Marcella Cornia, Silvia Cascianelli, Lorenzo Baraldi, and Rita Cucchiara. The unreasonable effectiveness of clip features for image captioning: an experimental analysis. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4662–4670, 2022. 4

[7] Samyadeep Basu, Nanxuan Zhao, Vlad I Morariu, Soheil Feizi, and Varun Manjunatha. Localizing and editing knowledge in text-to-image generative models. In *The Twelfth International Conference on Learning Representations*, 2023. 3

[8] Samyadeep Basu, Keivan Rezaei, Priyatham Kattakinda, Vlad I Morariu, Nanxuan Zhao, Ryan A Rossi, Varun Manjunatha, and Soheil Feizi. On mechanistic knowledge localization in text-to-image generative models. In *Forty-first International Conference on Machine Learning*, 2024. 3

[9] Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio Calmon, and Himabindu Lakkaraju. Interpreting clip with sparse linear concept embeddings (splice). *Advances in Neural Information Processing Systems*, 37:84298–84328, 2024. 4

[10] Jorge E Camargo and Fabio A González. Multimodal latent topic analysis for image collection summarization. *Information Sciences*, 328:270–287, 2016. 2

[11] George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4750–4759, 2022. 2

[12] George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A Efros, and Jun-Yan Zhu. Generalizing dataset distillation via deep generative prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3739–3748, 2023. 1, 2

[13] Jeffrey A Chan-Santiago, Praveen Tirupattur, Gaurav Kumar Nayak, Gaowen Liu, and Mubarak Shah. Mgd ˜3: Mode-guided dataset distillation using diffusion models. *arXiv preprint arXiv:2505.18963*, 2025. 1, 2, 5, 6, 7, 8, 9, 20, 21, 22, 23, 24, 25, 26, 27, 28

[14] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11472–11481, 2022. 3

[15] David J Crandall, Lars Backstrom, Daniel Huttenlocher, and Jon Kleinberg. Mapping the world's photos. In *Proceedings of the 18th international conference on World wide web*, pages 761–770, 2009. 1, 2

[16] Justin Cui, Ruochen Wang, Si Si, and Cho-Jui Hsieh. Scaling up dataset distillation to imagenet-1k with constant memory. In *International Conference on Machine Learning*, pages 6565–6590. PMLR, 2023. 2

[17] Giannis Daras and Alexandros G Dimakis. Multiresolution textual inversion. *arXiv preprint arXiv:2211.17115*, 2022. 3

[18] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5203–5212, 2020. 2

[19] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Mid-level visual element discovery as discriminative mode seeking. *Advances in neural information processing systems*, 26, 2013. 2

[20] Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei A Efros. What makes paris look like paris? *Communications of the ACM*, 58(12):103–110, 2015. 1, 2

[21] Luca Eyring, Shyamgopal Karthik, Karsten Roth, Alexey Dosovitskiy, and Zeynep Akata. Reno: Enhancing one-step text-to-image models through reward-based noise optimization. *Advances in Neural Information Processing Systems*, 37:125487–125519, 2024. 3

[22] Chao Feng, Ziyang Chen, Aleksander Holynski, Alexei A Efros, and Andrew Owens. Gps as a control signal for image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2766–2778, 2025. 3

[23] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023. 6

[24] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-

image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2, 5

[25] Yossi Gandelsman, Alexei A Efros, and Jacob Steinhardt. Interpreting clip's image representation via text-based decomposition. *arXiv preprint arXiv:2310.05916*, 2023. 4

[26] Shiry Ginosar, Kate Rakelly, Sarah Sachs, Brian Yin, and Alexei A Efros. A century of portraits: A visual historical record of american high school yearbooks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1–7, 2015. 2, 3

[27] Sonal Goel, Sarthak Ahuja, A V Subramanyam, and Ponnurangam Kumaraguru. # visualhashtags: Visual summarization of social media events using mid-level visual elements. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1434–1442, 2017. 2

[28] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2, 3

[29] Jianyang Gu, Saeed Vahidian, Vyacheslav Kungurtsev, Haonan Wang, Wei Jiang, Yang You, and Yiran Chen. Efficient dataset distillation via minimax diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15793–15803, 2024. 2, 1

[30] Xiefan Guo, Jinlin Liu, Miaomiao Cui, Jiankai Li, Hongyu Yang, and Di Huang. Initno: Boosting text-to-image diffusion models via initial noise optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9380–9389, 2024. 3

[31] Kimia Hamidieh, Haoran Zhang, Walter Gerych, Thomas Hartvigsen, and Marzyeh Ghassemi. Identifying implicit social biases in vision-language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 547–561, 2024. 8

[32] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3

[33] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[34] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 2, 5

[35] Drew A Hudson, Daniel Zoran, Mateusz Malinowski, Andrew K Lampinen, Andrew Jaegle, James L McClelland, Loic Matthey, Felix Hill, and Alexander Lerchner. Soda: Bottleneck diffusion models for representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23115–23127, 2024. 3

[36] Yong Jae Lee, Alexei A Efros, and Martial Hebert. Style-aware mid-level representation for discovering visual connections in space and time. In *Proceedings of the IEEE international conference on computer vision*, pages 1857–1864, 2013. 1, 2

[37] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2, 3

[38] Korrawe Karunratanakul, Konpat Preechakul, Emre Aksan, Thabo Beeler, Supasorn Suwajanakorn, and Siyu Tang. Optimizing diffusion noise can serve as universal motion priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1334–1345, 2024. 3

[39] Ira Kemelmacher-Shlizerman, Supasorn Suwajanakorn, and Steven M Seitz. Illumination-aware age progression. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3334–3341, 2014. 2

[40] Lyndon S Kennedy and Mor Naaman. Generating diverse and representative image search results for landmarks. In *Proceedings of the 17th international conference on World Wide Web*, pages 297–306, 2008. 1, 2

[41] Byungjun Kim, Soobin Um, and Jong Chul Ye. Diverse text-to-image generation via contrastive noise optimization. *arXiv preprint arXiv:2510.03813*, 2025. 3

[42] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[43] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2, 3

[44] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022. 2, 3, 4

[45] Erik G Learned-Miller. Data driven image models through continuous joint alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(2):236–250, 2005. 3

[46] Stefan Lee, Nicolas Maisonneuve, David Crandall, Alexei A Efros, and Josef Sivic. Linking past to present: Discovering style in two centuries of architecture. In *IEEE International Conference on Computational Photography*, 2015. 1, 2

[47] Saehyung Lee, Sanghyuk Chun, Sangwon Jung, Sangdoo Yun, and Sungroh Yoon. Dataset condensation with contrastive signals. In *International Conference on Machine Learning*, pages 12352–12364. PMLR, 2022. 2

[48] Yipeng Leng, Qiangjuan Huang, Zhiyuan Wang, Yangyang Liu, and Haoyu Zhang. Diffusegae: controllable and high-fidelity image manipulation from disentangled representation. In *Proceedings of the 5th ACM International Conference on Multimedia in Asia*, pages 1–7, 2023. 3

[49] Hang Li, Chengzhi Shen, Philip Torr, Volker Tresp, and Jindong Gu. Self-discovering interpretable diffusion latent directions for responsible text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12006–12016, 2024. 4

[50] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 2, 4

[51] Yao Li, Lingqiao Liu, Chunhua Shen, and Anton van den Hengel. Mining mid-level visual patterns with deep cnn

activations. *International Journal of Computer Vision*, 121 (3):344–364, 2017. 2

[52] Ziyi Li, Qinye Zhou, Xiaoyun Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Open-vocabulary object segmentation with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7667–7676, 2023. 3

[53] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 3

[54] Hanchao Liu, Xiaohang Zhan, Shaoli Huang, Tai-Jiang Mu, and Ying Shan. Programmable motion generation for open-set motion control tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1399–1408, 2024. 3

[55] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European conference on computer vision*, pages 423–439. Springer, 2022. 3

[56] Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Evaluating societal representations in diffusion models. *Advances in Neural Information Processing Systems*, 36:56338–56351, 2023. 2, 3, 4, 7, 1

[57] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *Advances in Neural Information Processing Systems*, 36:47500–47510, 2023. 3

[58] Yan Luo, Min Shi, Muhammad Osama Khan, Muhammad Muneeb Afzal, Hao Huang, Shuaihang Yuan, Yu Tian, Luo Song, Ava Kouhana, Tobias Elze, et al. Fairclip: Harnessing fairness in vision-language learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12289–12301, 2024. 8

[59] Chaofan Ma, Yuhuan Yang, Chen Ju, Fei Zhang, Jinxiang Liu, Yu Wang, Ya Zhang, and Yanfeng Wang. Diffusionseg: Adapting diffusion towards unsupervised object discovery. *arXiv preprint arXiv:2303.09813*, 2023. 3

[60] Benyuan Meng, Qianqian Xu, Zitai Wang, Xiaochun Cao, and Qingming Huang. Not all diffusion model activations have been evaluated as discriminative features. *Advances in Neural Information Processing Systems*, 37:55141–55177, 2024. 3

[61] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 6

[62] Erik G Miller, Nicholas E Matsakis, and Paul A Viola. Learning from one example through shared densities on transforms. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, pages 464–471. IEEE, 2000. 3

[63] Ryugo Morita, Stanislav Frolov, Brian Bernhard Moser, Takahiro Shirakawa, Ko Watanabe, Andreas Dengel, and Jinjia Zhou. Tkg-dm: Training-free chroma key content generation diffusion model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13031–13040, 2025. 7

[64] Koichi Namekata, Amirmojtaba Sabour, Sanja Fidler, and Seung Wook Kim. Emerdiff: Emerging pixel-level semantic knowledge in diffusion models. *arXiv preprint arXiv:2401.11739*, 2024. 3

[65] Dolev Ofri-Amar, Michal Geyer, Yoni Kasten, and Tali Dekel. Neural congealing: Aligning images to a joint semantic atlas. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19403–19412, 2023. 3

[66] Sakuya Ota, Qing Yu, Kent Fujiwara, Satoshi Ikehata, and Ikuro Sato. Pino: Person-interaction noise optimization for long-duration and customizable motion generation of arbitrary-sized groups. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10676–10685, 2025. 3

[67] Rishubh Parihar, Abhijnya Bhat, Abhipsa Basu, Saswat Mallick, Jogendra Nath Kundu, and R Venkatesh Babu. Balancing act: distribution-guided debiasing in diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6668–6678, 2024. 4

[68] Yong-Hyun Park, Mingi Kwon, Jaewoong Choi, Junghyo Jo, and Youngjung Uh. Understanding the latent space of diffusion models through the lens of riemannian geometry. *Advances in Neural Information Processing Systems*, 36: 24129–24142, 2023. 3

[69] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 2, 8, 1

[70] William Peebles, Jun-Yan Zhu, Richard Zhang, Antonio Torralba, Alexei A Efros, and Eli Shechtman. Gan-supervised dense visual alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13470–13481, 2022. 2, 3, 5, 6, 7, 9, 20, 21, 22, 23, 24, 25, 26, 27, 28

[71] Koutilya Pnvr, Bharat Singh, Pallabi Ghosh, Behjat Siddiquie, and David Jacobs. Ld-znet: A latent diffusion approach for text-based image segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4157–4168, 2023. 3

[72] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10619–10629, 2022. 3

[73] Zipeng Qi, Lichen Bai, Haoyi Xiong, and Zeke Xie. Not all noises are created equally: Diffusion noise selection and optimization. *arXiv preprint arXiv:2407.14041*, 2024. 3

[74] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2, 4, 6

[75] Konstantinos Rematas, Basura Fernando, Frank Dellaert, and Tinne Tuytelaars. Dataset fingerprints: Exploring image collections through data mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4867–4875, 2015. 2

[76] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 3

[77] Fawaz Sammani and Nikos Deligiannis. Interpreting and analysing clip's zero-shot image classification via mutual knowledge. *Advances in Neural Information Processing Systems*, 37:39597–39631, 2024. 4

[78] Dvir Samuel, Rami Ben-Ari, Matan Levy, Nir Darshan, and Gal Chechik. Where's waldo: Diffusion features for personalized segmentation and retrieval. *Advances in Neural Information Processing Systems*, 37:128160–128181, 2024. 3

[79] Dvir Samuel, Rami Ben-Ari, Simon Raviv, Nir Darshan, and Gal Chechik. Generating images of rare concepts using pre-trained diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4695–4703, 2024. 3

[80] Antonio Sclocchi, Alessandro Favero, and Matthieu Wyart. A phase transition in diffusion models reveals the hierarchical nature of data. *Proceedings of the National Academy of Sciences*, 122(1):e2408799121, 2025. 3

[81] SG_161222. Realistic vision v6.0 b1. https://civitai.com/models/4201/realistic-vision-v60-b1, 2023. Accessed: 2025-11-12. 5

[82] Shuai Shen, Wanhua Li, Xiaobing Wang, Dafeng Zhang, Zhezhu Jin, Jie Zhou, and Jiwen Lu. Clip-cluster: Clip-guided attribute hallucination for face clustering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 20786–20795, 2023. 4

[83] Yingdong Shi, Changming Li, Yifan Wang, Yongxiang Zhao, Anqi Pang, Sibei Yang, Jingyi Yu, and Kan Ren. Dissecting and mitigating diffusion bias via mechanistic interpretability. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8192–8202, 2025. 4

[84] Ioannis Siglidis, Aleksander Holynski, Alexei A Efros, Mathieu Aubry, and Shiry Ginosar. Diffusion models as data mining tools. In *European Conference on Computer Vision*, pages 393–409. Springer, 2024. 1, 2

[85] Ian Simon, Noah Snavely, and Steven M Seitz. Scene summarization for online image collections. In *2007 IEEE 11th International conference on computer vision*, pages 1–8. IEEE, 2007. 1, 2

[86] Saurabh Singh, Abhinav Gupta, and Alexei A. Efros. Unsupervised discovery of mid-level discriminative patches, 2012. 2

[87] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3, 4

[88] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 3

[89] Duo Su, Junjie Hou, Weizhi Gao, Yingjie Tian, and Bowen Tang. D^ 4: Dataset distillation via disentangled diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5809–5818, 2024. 1, 2, 5, 6, 7, 9, 20, 21, 22, 23, 24, 25, 26, 27, 28

[90] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. What makes tom hanks look like tom hanks. In *Proceedings of the IEEE international conference on computer vision*, pages 3952–3960, 2015. 2

[91] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36:1363–1389, 2023. 3

[92] Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, and Mar Gonzalez-Franco. Diffuse attend and segment: Unsupervised zero-shot segmentation using stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3554–3563, 2024. 3

[93] Kai Wang, Bo Zhao, Xiangyu Peng, Zheng Zhu, Shuo Yang, Shuo Wang, Guan Huang, Hakan Bilen, Xinchao Wang, and Yang You. Cafe: Learning to condense dataset by aligning features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12196–12205, 2022. 2

[94] Yingheng Wang, Yair Schiff, Aaron Gokaslan, Weishen Pan, Fei Wang, Christopher De Sa, and Volodymyr Kuleshov. Infodiffusion: Representation learning using information maximizing diffusion models. In *International conference on machine learning*, pages 36336–36354. PMLR, 2023. 3

[95] Yuxin Wen, Yuchen Liu, Chen Chen, and Lingjuan Lyu. Detecting, explaining, and mitigating memorization in diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. 7

[96] Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1206–1217, 2023. 3

[97] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023. 6

[98] Zeyuan Yin, Eric Xing, and Zhiqiang Shen. Squeeze, recover and relabel: Dataset condensation at imagenet scale from a new perspective. *Advances in Neural Information Processing Systems*, 36:73582–73603, 2023. 2

[99] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan

Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *Advances in Neural Information Processing Systems*, 36:45533–45547, 2023. 3

[100] Kaiwen Zhang, Yifan Zhou, Xudong Xu, Bo Dai, and Xingang Pan. Diffmorpher: Unleashing the capability of diffusion models for image morphing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7912–7921, 2024. 7

[101] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[102] Xulu Zhang, Xiao-Yong Wei, Jinlin Wu, Tianyi Zhang, Zhaoxiang Zhang, Zhen Lei, and Qing Li. Compositional inversion for stable diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7350–7358, 2024. 5

[103] Bo Zhao and Hakan Bilen. Dataset condensation with differentiable siamese augmentation. In *International Conference on Machine Learning*, pages 12674–12685. PMLR, 2021. 2

[104] Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6514–6523, 2023.

[105] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. *arXiv preprint arXiv:2006.05929*, 2020. 2

[106] Jun-Yan Zhu, Yong Jae Lee, and Alexei A Efros. Averageexplorer: Interactive exploration and alignment of visual data collections. *ACM Transactions on Graphics (TOG)*, 33(4):1–11, 2014. 2, 3

[107] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *European conference on computer vision*, pages 597–613. Springer, 2016. 2