ConceptBed: Evaluating Concept Learning Abilities of Text-to-Image Diffusion Models

Maitreya Patel[1]*, Tejas Gokhale[2], Chitta Baral[1], Yezhou Yang[1]

[1] Arizona State University
[2] University of Maryland Baltimore County

Abstract

The ability to understand visual concepts and replicate and compose these concepts from images is a central goal for computer vision. Recent advances in text-to-image (T2I) models have lead to high definition and realistic image quality generation by learning from large databases of images and their descriptions. However, the evaluation of T2I models has focused on photorealism and limited qualitative measures of visual understanding. To quantify the ability of T2I models in learning and synthesizing novel visual concepts (*a.k.a.* personalized T2I), we introduce ConceptBed, a large-scale dataset that consists of 284 unique visual concepts, and 33K composite text prompts. Along with the dataset, we propose an evaluation metric, Concept Confidence Deviation (`CCD`), that uses the confidence of oracle concept classifiers to measure the alignment between concepts generated by T2I generators and concepts contained in target images. We evaluate visual concepts that are either objects, attributes, or styles, and also evaluate four dimensions of compositionality: counting, attributes, relations, and actions. Our human study shows that `CCD` is highly correlated with human understanding of concepts. Our results point to a trade-off between learning the concepts and preserving the compositionality which existing approaches struggle to overcome. The data, code, and interactive demo is available at: https://conceptbed.github.io/

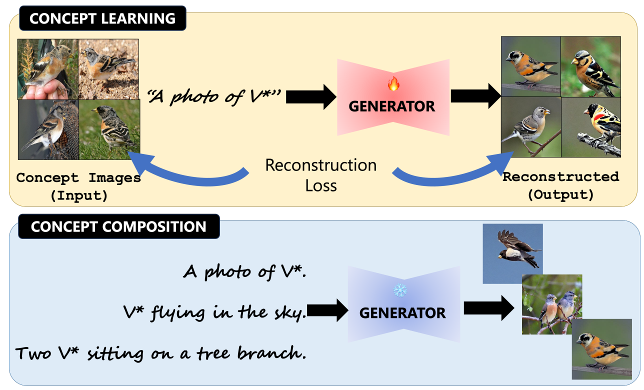

Figure 1: Visual concept learners such as textual inversion models learn to *"invert"* a set of images (about a concept c) into a text embedding V^*, and then use this learned textual concept in new text prompts to generate images of concept c under different contexts and by performing novel compositions with other concepts. The proposed ConceptBed dataset along with the evaluation metric `CCD` allows us to comprehensively and quantifiably evaluate concept learning abilities of text-to-image diffusion models.

1 Introduction

Humans reason about the visual world by aggregating entities that they see into "visual concepts": both `cats` and `elephants` are `animals`, and both `palms` and `pines` are `trees`. We use natural language to describe images and things that we see. Although this type of visual concept learning is well-defined in human psychology (Murphy 2004), it remains elusive in the context of data-driven techniques capable of learning and reasoning from images and their natural language descriptions.

Text-to-Image (T2I) generative models are trained to translate natural language phrases into images that correspond to that input. High-quality T2I models, therefore, serve as a link between human-level concepts (expressed in natural language) and their visual representations and are one way to reproduce visual concepts. On the other hand, this has also sparked interest in visual concept learning (*a.k.a.* personalized T2I) through the procedure of *"image inversion"* – to translate one or many images corresponding to a visual concept into a latent representation of that visual concept. While earlier methods primarily explored image inversion using generative adversarial networks (Xia et al. 2022), methods such as Textual Inversion (Gal et al. 2022) and Dreambooth (Ruiz et al. 2022) combine image inversion with T2I – this has led to an effective way to quickly learn concepts from a few images and reproduce them in novel combinations and compositions with other concepts, attributes, styles, etc. These methods aim to learn concepts with minimal reference images by fine-tuning pre-trained text-conditioned diffusion models (Figure 1). Therefore this paradigm of T2I and image inversion is a powerful new way of learning and reproducing concepts.

Within this paradigm of novel visual concept learning via image inversion, two primary evaluation criteria have emerged: (1) concept alignment, which assesses the corre-

*Corresponding Author: maitreya.patel@asu.edu

spondence between the generated images and the target concept images, and (2) composition alignment, which evaluates whether the generated images maintain compositionality. Previous studies have been small scale, evaluating only a small number of hand-picked concepts and compositions; as such making generic claims via such findings is difficult. Furthermore, the established evaluation metrics such as DINO-based cosine similarity (Ruiz et al. 2022) (for measuring concept alignment), KID (Kumari et al. 2022) (for measuring the amount of concept overfitting), and CLIP-Score (Hessel et al. 2021) (for evaluating compositionality), have encountered challenges in accurately capturing human preferences. Consequently, there is a growing need for better automated evaluations.

Therefore, we introduce CONCEPTBED, comprehensive dataset and evaluation framework that is aligned with human preferences. The CONCEPTBED dataset comprises 284 distinct concepts and approximately 33,000 composite text prompts, which can be further extended using the provided automatic realistic dataset creation pipeline. The dataset focuses on four diverse concept learning evaluation tasks: learning styles, learning objects, learning attributes, and compositional reasoning. To gain a deeper understanding of previous methodologies, we incorporate four composition categories – action, attribution, counting, and relations.

We use our large-scale dataset to evaluate concept learners, by developing a novel evaluation metric called Concept Confidence Deviation (CCD). We conduct a human study and find that relative evaluations of models in terms of CCD are well aligned with human preferences. Therefore, CCD combined with the CONCEPTBED dataset, offers an alternative to existing evaluation strategies, facilitating more effective large-scale evaluations. For each evaluation criteria, we train supervised classifiers (oracles) to detect whether generated concept images are accurate. Subsequently, the confidence scores from these oracles are utilized to calculate the instance-level concept deviations of the generated concept images in relation to the reference target ground truth images using the proposed CCD metric. This approach enables us to assess concept and composition alignment more effectively. We further show that CCD calculated using a pretrained few-shot classifier also maintains a high correlation with human preferences. This allows CCD to measure concept alignment on unseen concepts.

We conduct extensive experiments on four recently proposed concept learning methodologies. In total, we fine-tune approximately 1100 models (one model per concept) and generate over 500,000 concept-specific images. Our results reveal a surprising trade-off between concept alignment and composition alignment, wherein methods excelling at concept alignment tend to fall short in preserving compositions and vice versa. This suggests that previous concept learning approaches are either highly overfitted or severely underfitted. Furthermore, our experiments demonstrate that utilizing a pre-trained CLIP (Radford et al. 2021) textual encoder aids in maintaining compositionality, but it lacks the flexibility required to learn complex concepts, such as `sketch`.

In summary, we make the following **key contributions:**

- We introduce CONCEPTBED, a comprehensive benchmark for grounded quantitative evaluations of text-conditioned concept learners.
- The Concept Confidence Deviation (CCD) evaluation metric, measures the learners' ability to preserve concepts and compositions. We demonstrate a strong correlation between CCD and human preferences.
- Through extensive experiments with 1,100+ models, we identify shortcomings in prior works and suggest future research directions. CONCEPTBED sets a standard for evaluating personalized text-to-image generative models.

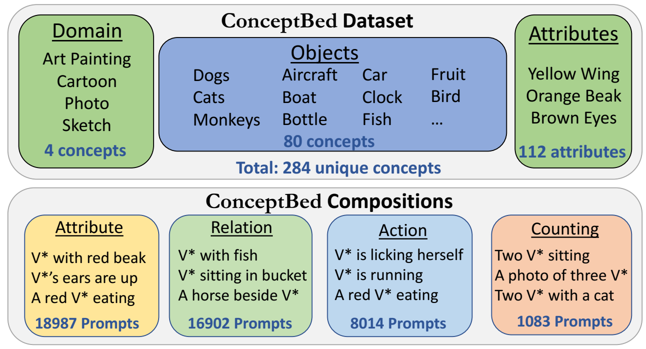

Figure 2: A summary of the CONCEPTBED dataset for large-scale grounded evaluations of concept learners. The collection of concepts is categorized into three classes: (1) Domain, (2) Objects, and (3) Attributes. CONCEPTBED has 284 unique concepts and four compositional categories. Here, V* is a learned concept.

2 Preliminaries

Prior studies on concept learning have focused on text-conditioned diffusion models, such as Textual Inversion (Gal et al. 2022), DreamBooth (Ruiz et al. 2022), and Custom Diffusion (Kumari et al. 2022). These models operate within the T2I paradigm, where a text prompt (y) serves as input to generate the corresponding image (x_{gen}) representing the given prompt y. A popular approach within T2I is the Latent Diffusion Model (LDM) (Rombach et al. 2022), which incorporates two key modules:

1. Textual Encoder (C_θ): This module generates embeddings corresponding to the input text prompt;
2. Generator (ϵ_ϕ): The generator estimates the noise iteratively from the input randomly sampled matrix at timestamp t (z_t), conditioned on the text.

Since T2I models solely consider text input, the target concept (c) is represented in terms of text tokens. These tokens can subsequently be employed to generate images associated with concept c. Therefore, in Textual Inversion, the concept learning task is approached as an image inversion problem, aiming to map the target concept back to the text-embedding space.

Let V^* denote the text tokens corresponding to the learned concept c. Once the optimal mapping from V^* to the target concept is determined, we can generate concept-specific images using the LDM by providing V^* in the text prompt.

Algorithm 1: Concept Confidence Deviation

Input: Concept fine-tuned models $G \in \{g_c\}$, $c \in C_{\text{ConceptBed}}$;
Oracles $F_t \in \{F_{PAC}, F_{Imagenet}, F_{CUBS}, F_{VQA}\}$;
Reference set of concept images $X^{ref} \in \{x_c\}$;
Target set of prompts $Y \in \{y_c\}$;
Output: Estimated CCD

Initialize: $score = []$; $p^{real} = []$
for $c \in C_{\text{ConceptBed}}$ **do**
  $p^{real} = []$
  if $t = VQA$ **then**
    $c = 3$
  for $x = 1 \ldots M$ **do**
    $p^{real} \leftarrow F_t(x_i, c)$
  $\bar{p}^{real} = \frac{1}{M}\sum_{i=1}^{M} p_i^{real}$
  for $n = 1 \ldots N$ **do**
    $x_{gen} = g_c(y_c)$
    $score \leftarrow -1 * (F(x_{gen}, c) - \bar{p}^{real})$ {// Eq. 3}
CCD $= \frac{1}{NC}\sum_{i=1}^{NC} score_i$

Suppose we are provided with m images ($X_{1:m}$) of the target concept c. Now, in order to learn the text tokens V^* corresponding to the concept c from the set of images $X_{1:m}$, the Textual Inversion methodology aims to optimize V^* by reconstructing $X_{1:m}$ using the objective function of the LDM with frozen parameters θ and ϕ:

$$V^* = \underset{v}{\text{argmin}} \underset{\substack{x \in X_{1:m},\ t, \\ \epsilon \sim \mathcal{N}(0,1),\ z \sim \mathcal{E}(x)}}{\mathbb{E}} ||\epsilon - \epsilon_\phi(z_t, t, x, C_\theta(y))||_2^2 \quad (1)$$

In the case of DreamBooth and Custom Diffusion, instead of finding the optimal V^*, it optimizes the model parameter ϕ associated with the noise estimator (ϵ_ϕ). This optimization process enables the model to learn the mapping between randomly initialized V^* and the target concept c.

$$\phi^* = \underset{\phi}{\text{argmin}} \underset{\substack{x \in X_{1:m},\ t, \\ \epsilon \sim \mathcal{N}(0,1),\ z \sim \mathcal{E}(x)}}{\mathbb{E}} ||\epsilon - \epsilon_\theta(z_t, t, x, C_\phi(y))||_2^2 \quad (2)$$

Once ϕ^* is obtained, it can be used to generate images related to the target concept.[1]

Once the images are generated, in order to evaluate these generated images, it is essential to verify whether they align with the learned concepts while maintaining compositionality.

3 ConceptBed

In this section, we introduce ConceptBed, a comprehensive collection of concepts, designed to accurately estimate concept and composition alignment by quantifying deviations in the generated images. Later, we introduce the novel evaluation framework associated with ConceptBed. [2]

[1]DreamBooth and Custom-Diffusion use additional regularizer to improve compositionally by using same objective function on a diverse set of image-caption pairs.

[2]Please refer to the arxiv release for additional insights on the proposed dataset and evaluation framework: https://arxiv.org/abs/2306.04695.

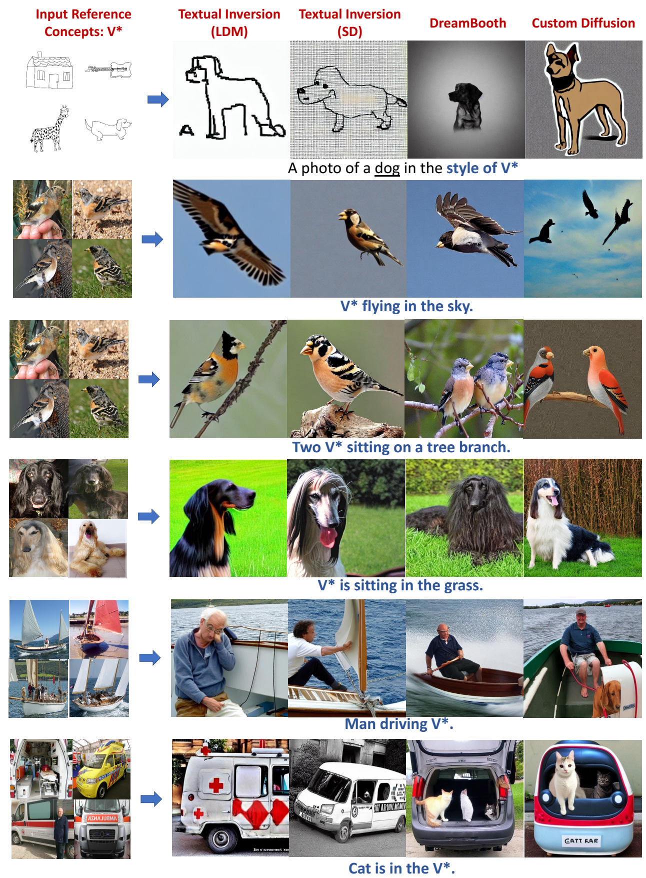

Figure 3: Qualitative examples showcasing the effectiveness of concept learners on the ConceptBed dataset. The leftmost column displays four instances of ground truth target concept images (V^*). Subsequent columns exhibit target concept-specific images generated by all baseline methods.

3.1 ConceptBed: Dataset Construction

ConceptBed incorporates existing datasets such as ImageNet (Deng et al. 2009), PACS (Li et al. 2017), CUB (Wah et al. 2011), and Visual Genome (Krishna et al. 2017), enabling the creation of a labeled dataset. Figure 2 provides an overview of the ConceptBed dataset.

Learning Styles. We use styles from the PACS dataset: `Art Painting`, `Cartoon`, `Photo`, and `Sketch`. Each style contains images corresponding to seven categories. The concept learner aims to use examples from one style as a reference and generate style-specific images for all seven entities.

Learning Objects. Extracting object-level concepts is accomplished through the utilization of the ImageNet dataset. It comprises 1000 low-level concepts from the WordNet (Fellbaum 2010) hierarchy. However, due to the presence of noise in ImageNet images and the lack of relevance to daily life for many concepts, we employ an automated filtering pipeline to ensure the usefulness and quality of the reference concept images. The pipeline involves extracting a list of low-level concepts and their parent concepts from

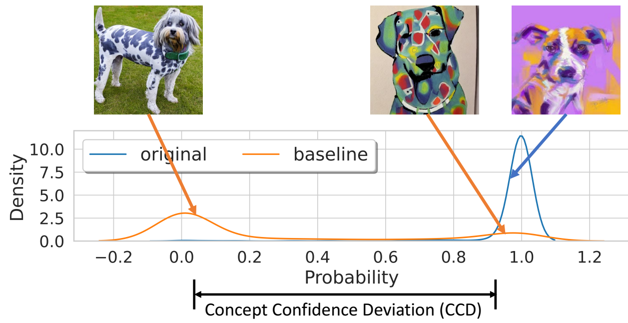

Figure 4: Intuitive illustration of the Concept Confidence Deviation (CCD) for the concept `Art Painting`. Blue and Orange are the probability distributions of the real and generated concept images.

ImageNet, followed by extracting text phrases from Visual Genome containing the concept as a subject in the caption. If an insufficient number of such captions exists (less than 10 in Visual Genome) or they cannot be found, the concepts are discarded. This filtering process results in 80 concepts such as (`brambling`, `squirrel monkey`, etc.). We select the top 100 high-quality images for each concept that will be used to train the concept learning methodologies.

Learning Attributes. Since ImageNet dataset images are not labeled based on the attributes present in the image, it is necessary to rely on datasets that provide attribute-level grounded labels. Therefore, we additionally employ the CUB dataset, which offers attribute-level labels (such as `orange wing`, `blue forehead`, etc.), enabling the ConceptBed to perform evaluations and measure the attribute-level performance of concept learners.

Compositional Reasoning. In addition to learning new concepts, it is crucial to maintain prior knowledge and associate the acquired concepts with it. To conduct these evaluations holistically, we use Visual Genome to extract captions in which the concept appears as the subject of the sentence. These captions are categorized into four composition categories (actions, attributes, counting, and relation) through few-shot classification using GPT3 (Brown et al. 2020). This categorization allows us to measure the performance of the baselines on each category, and an in-depth understanding of the varying difficulty levels of different compositions.

3.2 ConceptBed: Dataset Statistics

The ConceptBed dataset consists of 284 unique concepts, comprising 80 concepts from ImageNet, 200 concepts from CUB, and 4 concepts from PACS. In total, the dataset contains approximately 33,000 composite prompts for the evaluation of all 80 processed concepts from ImageNet, with each composite prompt having up to two composition categories. Out of these composite prompts, 18987, 16902, 8014, and 1083 prompts contribute to the attribute, relation, action, and counting categories, respectively.

Our dataset curation pipeline is flexible to be extended to larger datasets such as OpenImages-v7 (Kuznetsova et al. 2020) and LAION-5B (Schuhmann et al. 2022). However, it is important to note that this extension would significantly increase the resource requirements. With the introduction of this dataset, our primary objective is to provide a standardized and benchmarked evaluation framework for concept learners, enhancing research in the field.

3.3 CCD: Concept Confidence Deviation

Problem Statement. Consider a pre-trained text-conditioned diffusion model $g(\cdot)$, which can be further fine-tuned on a specific concept c such that $c \in \mathcal{C}_{\text{ConceptBed}}$. We assume the availability of concept-specific target images from the ConceptBed dataset, denoted as $\mathcal{D}_c^{real} \in \mathcal{D}^{test}_{\text{ConceptBed}}$. Denote the concept learner $g(\cdot)$ fine-tuned on concept c using $\mathcal{D}_c^{real}$ as $g_c(\cdot)$. First, we generate a collection of N images using the learned concept c, and denote this set of images as $\mathcal{D}_c^{gen} = \{x_i^{gen} = g_c(p_c^i, s^i);\ \forall i \in [0, N]\}$, where p_c^i is the concept-specific prompt and s is the random seed.

The alignment between two distributions (i.e., D_c^{real} and D_c^{gen}) is typically computed by first extracting features from the model m (i.e., $f_{real} = m(D^{real})$; $f_{gen} = m(D^{gen})$) and then employing a distance metric d (i.e., $score = d(f_{real}, f_{gen})$). Several combinations of models (m) and distance measures (d) have been used in prior work. For concept alignment, Ruiz et al. (2022) use m=DINO with d=Cos and Kumari et al. (2022) use m=Inception with d=KID. For composition alignment, all prior work utilizes m=CLIP with d=Cos. However, these methods fail to accurately capture the concept deviations within the generated images; rendering them ineffective in comparing performance across the methodologies (as shown in Section 4.2).

Concept Confidence Deviation (CCD). To address the above limitations, we propose training the oracle classifier F, specifically for the concept detection task using the ConceptBed training dataset, $\mathcal{D}^{train}_{\text{ConceptBed}}$. Then one can simply use $m = F$ and $d =$ Accuracy to verify whether x_{gen} is aligned with x_{real}. However, measuring accuracy does not allow instance-level evaluations. By leveraging the output probabilities of the oracle (concerning the concept label y_c), we can estimate the deviations associated with each generated image x_{gen} w.r.t. the output probabilities of real target images x_{real}. Concept Confidence Deviation is defined as:

$$\text{CCD} = -\mathop{\mathbb{E}}_{c}\left[\mathop{\mathbb{E}}_{x_{gen}}\left[F(y_c|x_{gen}) - \mathop{\mathbb{E}}_{x_{real}} F(y_c|x_{real})\right]\right]. \quad (3)$$

CCD first calculates the mean target probability on the test ground truth images and then measures the difference in probability of the generated images. CCD with negative or close to 0.0 values indicates that the generated images closely follow the distribution of the ground truth concept images. A positive CCD value suggests that the generated images deviate from the original distribution. Figure 4 shows an intuitive example of CCD by calculating the distance between two probability densities corresponding to the real and generated target concept.

Model	Concepts		Fine-grained$_{\text{CUB}}$		Composition
	Domain$_{\text{PACS}}$	Objects$_{\text{ImageNet}}$	Object-level	Attribute-level	
TI (LDM)	**0.0478**	0.0955	0.2289	0.1174	0.1906
TI (SD)	0.2456	**0.0472**	**0.0859**	**0.0332**	**0.1090**
DB	<u>0.6825</u>	0.0678	0.0963	0.0469	0.3527
CD	0.6206	<u>0.2085</u>	<u>0.3934</u>	<u>0.1743</u>	<u>0.4916</u>
Original	0.0000	0.0000	0.0000	0.000	0.0000

Table 1: Results of Concept Alignment Evaluation. The table shows the performance of concept learners evaluated using the `CCD` (↓) metric for Concepts (Domain$_{\text{PACS}}$, and Object$_{\text{ImageNet}}$), Fine-grained$_{\text{CUB}}$ (Object-level, and Attribute-level), and Composition. The best and worst performing models are indicated by bold and underlined numbers, respectively.

Models	Relation			Action			Attribute			Counting		
	CLIP	VQA.	CCD	CLIP	VQA	CCD	CLIP	VQA.	CCD	CLIP	VQA.	CCD
TI (LDM)	0.6589	66.60%	0.2074	0.6523	68.69%	0.2098	0.6599	72.22%	0.1331	0.6515	65.78%	0.1231
TI (SD)	0.6294	70.09%	0.1735	0.6274	70.81%	0.1884	<u>0.6360</u>	74.75%	0.1091	0.6301	68.38%	0.1020
DB	<u>0.7051</u>	82.20%	0.0542	<u>0.6995</u>	84.61%	0.0496	<u>0.6862</u>	82.24%	0.0355	**<u>0.6924</u>**	<u>78.90%</u>	<u>-0.0016</u>
CD	**<u>0.7065</u>**	**82.94%**	**0.0471**	**<u>0.7053</u>**	**86.35%**	**0.0347**	**<u>0.6940</u>**	**84.20%**	**0.0163**	<u>0.6921</u>	**<u>79.36%</u>**	**<u>-0.0054</u>**
SD	<u>0.7222</u>	83.42%	0.0403	<u>0.7178</u>	87.39%	0.0256	<u>0.7053</u>	83.85%	0.0184	<u>0.7085</u>	<u>81.07%</u>	<u>-0.0206</u>
Original	0.6626	87.45%	0.0000	0.6831	89.78%	0.0000	0.6306	85.79%	0.0000	0.6553	78.32%	0.0000

Table 2: Compositional Reasoning Evaluation Results. The table shows the performance of the prior works for Composition Alignment. CLIP (↑) is the traditional image-text alignment metric. VQA (↑) is the accuracy of the ViLT VQA classifier on generated boolean questions. And `CCD` (↓) is the composition deviations reported from the ViLT model with respect to its performance on original images. The best-performing model is indicated by bold numbers, while the performance that is higher than the original data is reported with underline.

3.4 Task Specific Evaluation Settings

To efficiently leverage the CONCEPTBED evaluation pipeline, we trained separate oracles on the corresponding CONCEPTBED datasets. Two different types of evaluations are conducted, each with its respective set of oracles: 1) concept alignment, measured by concept classifiers, and 2) compositional reasoning, measured by a VQA model.

Concept Alignment: Concept alignment evaluation was performed on all tasks, including the generated concept images with different composite text prompts. To evaluate the style, a ResNet18 (He et al. 2015) model is trained to distinguish the images between four style concepts. To evaluate the objects, a ConvNeXt (Liu et al. 2022) model is fine-tuned on 80 classes from the CONCEPTBED using the ImageNet training subset. The Concept Embedding Model (CEM) (Zarlenga et al. 2022) was trained on CUB to detect the concepts and attributes. Images corresponding to the concepts were generated for each task by following the prompts: "A photo of V*" for objects and "A photo of a <*entity-name*> in the style of V*" for styles. Here, <*entity-name*> belongs to the seven classes from PACS. The remaining task, composition, utilizes the same pre-trained ConvNeXt model for concept alignment, as CONCEPTBED compositions are specifically for 80 ImageNet concepts.

Compositional Reasoning: To measure the image-text alignment with respect to the input prompts, the concept-specific token (V*) was removed and replaced with the corresponding ground truth label (i.e., `dogs`, `cats`, etc.). The image-text similarity was then measured. Unlike previous works, CLIP was not used due to its inability to capture compositions (Thrush et al. 2022). Instead, taking after (Cho, Zala, and Bansal 2022), we propose to use a pre-trained ViLT (Kim, Son, and Kim 2021) as a VQA model for composition evaluations. Specifically, from each composite prompt, the boolean questions with positive answers are generated (Banerjee et al. 2021). As ViLT is essentially a classifier, the `CCD` can be calculated with respect to the confidence of the model associated with a "yes" answer.

4 Experiments & Results

In this section, we benchmark four state-of-the-art concept learning methodologies. We first explain the experimental setup and report the evaluation results using the CONCEPTBED framework along with human preferences. [3]

4.1 Experimental Setup

In our experiments, we study four text-conditioned diffusion modeling-based concept learning strategies: Textual Inversion (TI) on LDM and SD, DreamBooth (DB) (Ruiz et al. 2022), and Custom Diffusion (CD) (Kumari et al. 2022). We generate $N = 100$ images for all concepts to measure the concept alignment and $N = 3$ images for 33K composite text prompts. For a total of 284 concepts, we train all four baselines. This leads to 1100+ concept-specific fine-tuned models and we generate a total of $500,000$ images for eval-

[3]The arxiv release contains additional details about the experimental setup, results, and human evaluations: https://arxiv.org/abs/2306.04695

Models	Domain$_{\text{PACS}}$				Objects$_{\text{ImageNet}}$				Compositional Reasoning		
	DINO (↑)	KID (↓)	CCD (↓)	H.S. (↑)	DINO (↑)	KID (↓)	CCD (↓)	H.S. (↑)	CLIP (↑)	CCD (↓)	H.S. (↑)
TI (LDM)	0.5073	0.0117	0.0478	4.028	0.4708	0.0552	0.0955	4.069	0.6611	0.1684	2.851
TI (SD)	0.4104	0.0422	0.2456	4.084	0.4457	0.0294	0.0472	4.159	0.6309	0.1432	3.694
DB	0.3925	0.1101	0.6825	3.083	0.4525	0.0290	0.0678	4.075	0.6919	0.0344	3.556
CD	0.3956	0.0593	0.6206	3.164	0.4450	0.0492	0.2085	3.803	0.6968	0.0232	4.178
Correlation	0.6557	<u>-0.8252</u>	**-0.9515**	1.000	0.2787	<u>-0.5347</u>	**-0.9892**	1.000	<u>0.3486</u>	**-0.7342**	1.000

Table 3: Human Evaluations. Comparison of prior quantitative metrics and CCD metric with Human evaluations. DINO based pairwise cosine similarity is the prior evaluation metric (Ruiz et al. 2022). KID was used to measure the overfitting by (Kumari et al. 2022). CLIP (CLIPScore) is the traditional reference-free image-text similarity metric. CCD is our presented concept deviation-aware evaluation metric. H.S. denotes the corresponding Human Score. Here, Domain$_{\text{PACS}}$ and Object$_{\text{ImageNet}}$ evaluations are for concept alignment and composition alignment is for image-text similarity. A high negative correlation between CCD and human ratings implies strong alignment, as lower CCD and higher human ratings correspond to better performance.

Model	PACS Domain	ImageNet Object	ImageNet Composition
TI (LDM)	72.84	64.53	58.28
TI (SD)	52.25	70.79	65.42
DB	24.71	67.45	39.42
CD	20.12	52.06	26.31

Table 4: Recall. Percentage of generated images highly aligned (CCD $<= 0.0$) with the target concept images.

uations. To show the stability of CCD, we report the mean performance across the three seeds of oracle training.

4.2 Results

Concept Alignment. Table 1 shows the overall performance of the baselines in terms of CCD, where lower score indicates better performance. First, we can observe that CCD for *concept alignment* is low for the original images; suggesting that the oracle is certain about its predictions. Second, it can be inferred that Custom Diffusion performs poorly, while Textual Inversion (SD) outperforms the other methodologies except for the case of the learning styles. We attribute this behavior to differences in textual encoders. LDM trains the BERT-style textual encoder from scratch while SD uses pre-trained CLIP to condition the diffusion model. CLIP contains vast image-text knowledge leading to better performance on learning objects but less flexibility to learn different styles as a concept. Surprisingly, if we compare the *concept alignment* performance with and without composite prompts, we observe that the performance further drops significantly for all baseline methodologies when composite prompts are used. This shows that existing concept learning methodologies find it difficult to maintain the concepts whenever the prompt contains the composition.

Compositional Reasoning. Previously, we discussed concept alignment on composite prompts. Table 2 summarizes the evaluations on composition tasks. Here, we observe the complete opposite trend in results. Custom Diffusion outperforms the other approaches across the composition categories. This result shows the trade-off between learning concepts and at the same time maintaining compositionality in recent concept learning methodologies. Moreover, CLIPScore estimates the better performance of the baselines compared to the original image-text pairs which are inaccurate.

Qualitative Results. Figure 3 provides the qualitative examples of the concept learning. It can be inferred that Textual Inversion (LDM) learns the `sketch` concept very well (the first row), while DreamBooth and Custom Diffusion struggle to learn it. All baselines perform comparatively well in reproducing the learned concept (the second row). Interestingly, in the case of compositions, DreamBooth and Custom Diffusion perform well with the cost of losing the concept alignment (the last two rows). At the same time, textual inversion approaches cannot reproduce the compositions (like, *"Two V^*"*) but they maintain concept alignment. Overall, these qualitative examples align with our quantitative results and strengthen our evaluation framework.

Human Evaluations. We perform Human Evaluations using Amazon Mechanical Turk for both types of evaluations: 1) concept alignment – to measure the alignment between generated images and ground truth reference images on Domain$_{\text{PACS}}$ and Object$_{\text{ImageNet}}$, and 2) compositional reasoning – to measure the image-text alignment. For concept alignment, we ask human annotators to rate the likelihood of the target image the same as three reference images. While for compositional reasoning we simply ask the annotators to rate the likelihood alignment of the image and the corresponding caption. Table 3 summarizes the performance of prior and proposed (CCD) quantitative metrics *w.r.t.* the Human Score. KID performs better for domains than objects as image dynamics varies a lot in domains. (Kumari et al. 2022) proposed to use KID with LAION-retrieved concept images as a reference instead of ground truth due to the scarcity of reference images. However, CONCEPTBED alleviates this limitation. Therefore, we use actual ground truth images to report KID which is more accurate. It can be inferred that the CCD is strongly correlated with human preferences and outperforms the prior evaluation metrics by a large amount.

Percentage of highly aligned instances. Using CCD, we can further measure the recall of the concept learning models. DINO and KID metrics do not allow us to measure the recall. Hence, it becomes hard to investigate the actual

Models	ConvNeXt	Inception	ViT	Few-Shot
TI (LDM)	0.0955	0.0773	0.1165	0.0823
TI (SD)	0.0472	0.0201	0.0599	0.0489
DB	0.0678	0.0485	0.0786	0.0596
CD	0.2085	0.1845	0.2286	0.1384
Correlation	-0.9892	-0.9888	-0.9816	-0.9763

Table 5: Ablation. Effect of different oracle models to measure concept alignment using CCD.

quality of the generated images. Table 4 shows the recall ($\frac{\text{sample with CCD}<=0.0}{\text{total samples}} * 100$) for the concept alignment shown in Table 1. It can be inferred that Custom-Diffusion can work once in every four generation attempts. While Textual Inversion will work at least once in every two attempts. At the same time, when composition prompts are provided, Textual Inversion consistently maintains the concept alignment at the cost of achieving the composition alignment.

Generalization. Fine-tuned oracles cannot be generalized to unseen concepts; making CCD unreliable on OOD concepts. Hence, we propose to utilize a few-shot classifier (5-way 5-shot) instead, which can allow the generalization to unseen concepts while maintaining a high correlation (shown in Table 5). This shows the effectiveness of using confidence and CCD as the alternative to the DINO, KID, and CLIP.

5 Related Work

Concept Learning. Concept learning encompasses various problem statements and approaches, depending on the perspective adopted. Concept Bottleneck Models (CBMs) (Koh et al. 2020) and Concept Embedding Models (CEMs) (Zarlenga et al. 2022) treat object attributes as concepts and propose classification strategies to identify these concepts. Neuro Symbolic Concept Learner (NS-CL) (Mao et al. 2019) aims to learn visual concepts by associating them with language semantics, enabling the model to perform visual question answering. Image Inversion Style Concept Learning (Xia et al. 2022), takes a different approach. Its objective is to invert a given concept image back into the latent space of a pre-trained model. However, text-based concept composition is not possible for such models.

Text-to-Image Generative Models. With advances in vector quantization (Van Den Oord, Vinyals et al. 2017) and diffusion modeling (Rombach et al. 2022), text-to-image generation has improved its performance. Notable works such as DALL-E (Ramesh et al. 2021) train transformer models. While current state-of-the-art, diffusion-based text-to-image models such as GLIDE (Nichol et al. 2022), LDM (Rombach et al. 2022), and Imagen (Saharia et al. 2022), have surpassed prior approaches (such as StackGAN (Zhang et al. 2017), StackGAN++ (Zhang et al. 2018), TReCS (Koh et al. 2021), and DALL-E (Ramesh et al. 2021)) and achieved superior performance. Pixart-α (Chen et al. 2023) and ECLIPSE (Patel et al. 2023) further enhances T2I methods without depending on heavy compute. Additionally, as shown by (Saxon and Wang 2023), these T2I models also have multilingual concept understanding to a certain extent.

Text-to-Image Concept Learning. Text-conditioned diffusion models, such as LDM, have demonstrated their potential for learning novel visual concepts with only a few reference images. Textual Inversion (Gal et al. 2022) proposes learning the embedding corresponding to the placeholder (V^*) through optimization. DreamBooth (Ruiz et al. 2022) suggests optimizing the UNet parameters instead of optimizing the placeholder embedding. Custom Diffusion (Kumari et al. 2022) combines both approaches by optimizing the placeholder and key/value weights from the cross-attention layers for faster concept learning. These concept learners are essentially text-conditioned diffusion models and inherit the same limitations of diffusion models. One limitation is the overfitting of concepts and language drift. By optimizing model parameters on a handful of reference images, it is highly likely that the model might overfit the given concept and cannot maintain compositionality. Therefore, in this paper, we propose ConceptBed for systematic evaluations.

Text-to-Image Generative Model Evaluations. Evaluating generative models is not widely studied. The FID (Heusel et al. 2017) score is commonly used to measure generated image quality. CLIPScore (Hessel et al. 2021) is another popular evaluation metric for reference-free image-text alignment. Another study focuses on compositional evaluations of text-to-image models on small subsets (CU-Birds and Oxford-Flowers) (Park et al. 2021). DALL-Eval (Cho, Zala, and Bansal 2022) evaluates reasoning skills on synthetic datasets and social biases of text-to-image generative models. DALL-Eval, VISOR (Gokhale et al. 2022), LAYOUTBENCH (Cho et al. 2023) evaluates spatial reasoning abilities. Parallel work T2I CompBench (Huang et al. 2023) also adopts the idea of VQA for accurate composition evaluations. Although text-to-image model evaluations are well-explored, they lack concept-specific assessments and cannot be used for evaluating concept learning. Therefore, ConceptBed attempts to overcome this gap in evaluations of novel visual concept learning abilities.

6 Conclusion

In this paper, we introduce a novel benchmark called ConceptBed designed to assess the efficacy of text-conditioned diffusion models in learning new concepts (*a.k.a.* personalized T2I). The ConceptBed benchmark encompasses an end-to-end evaluation pipeline, a comprehensive concept library, and a novel Concept Confidence Deviation (CCD) evaluation metric. We conduct evaluations based on two key criteria: concept alignment and composition alignment. Through extensive experiments, we demonstrate that existing text-conditioned diffusion model-based concept learners exhibit significant limitations in their performance. We perform human evaluations to validate the effectiveness of our proposed evaluation metric (CCD), which showcases a strong correlation with human preferences. This finding positions CCD as a viable alternative to human judgments, enabling large-scale and comprehensive evaluations. ConceptBed represents the first large-scale concept-learning dataset that facilitates precise and accurate evaluations of personalized text-to-image generative models.

Acknowledgments

This work was supported by NSF RI grants #1750082 and #2132724, and a grant from Meta AI Learning Alliance. The views and opinions of the authors expressed herein do not necessarily state or reflect those of the funding agencies and employers.

References

Banerjee, P.; Gokhale, T.; Yang, Y.; and Baral, C. 2021. WeaQA: Weak Supervision via Captions for Visual Question Answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 3420–3435.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chen, J.; Yu, J.; Ge, C.; Yao, L.; Xie, E.; Wu, Y.; Wang, Z.; Kwok, J.; Luo, P.; Lu, H.; et al. 2023. PixArt-alpha: Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis. *arXiv preprint arXiv:2310.00426*.

Cho, J.; Li, L.; Yang, Z.; Gan, Z.; Wang, L.; and Bansal, M. 2023. Diagnostic Benchmark and Iterative Inpainting for Layout-Guided Image Generation. *arXiv preprint arXiv:2304.06671*.

Cho, J.; Zala, A.; and Bansal, M. 2022. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. *arXiv preprint arXiv:2202.04053*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Fellbaum, C. 2010. WordNet. In *Theory and applications of ontology: computer applications*, 231–243. Springer.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Gokhale, T.; Palangi, H.; Nushi, B.; Vineet, V.; Horvitz, E.; Kamar, E.; Baral, C.; and Yang, Y. 2022. Benchmarking Spatial Relationships in Text-to-Image Generation. *arXiv preprint arXiv:2212.10015*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep residual learning for image recognition. arXiv 2015. *arXiv preprint arXiv:1512.03385*, 14.

Hessel, J.; Holtzman, A.; Forbes, M.; Le Bras, R.; and Choi, Y. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 7514–7528.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Huang, K.; Sun, K.; Xie, E.; Li, Z.; and Liu, X. 2023. T2I-CompBench: A Comprehensive Benchmark for Open-world Compositional Text-to-image Generation. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Kim, W.; Son, B.; and Kim, I. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, 5583–5594. PMLR.

Koh, J. Y.; Baldridge, J.; Lee, H.; and Yang, Y. 2021. Text-to-image generation grounded by fine-grained user attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 237–246.

Koh, P. W.; Nguyen, T.; Tang, Y. S.; Mussmann, S.; Pierson, E.; Kim, B.; and Liang, P. 2020. Concept bottleneck models. In *International Conference on Machine Learning*, 5338–5348. PMLR.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123: 32–73.

Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J.-Y. 2022. Multi-Concept Customization of Text-to-Image Diffusion. *arXiv preprint arXiv:2212.04488*.

Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.; Kolesnikov, A.; et al. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7): 1956–1981.

Li, D.; Yang, Y.; Song, Y.-Z.; and Hospedales, T. M. 2017. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, 5542–5550.

Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; and Xie, S. 2022. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11976–11986.

Mao, J.; Gan, C.; Kohli, P.; Tenenbaum, J. B.; and Wu, J. 2019. The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision. In *International Conference on Learning Representations*.

Murphy, G. 2004. *The big book of concepts*. MIT press.

Nichol, A. Q.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; Mcgrew, B.; Sutskever, I.; and Chen, M. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *International Conference on Machine Learning*, 16784–16804. PMLR.

Park, D. H.; Azadi, S.; Liu, X.; Darrell, T.; and Rohrbach, A. 2021. Benchmark for compositional text-to-image synthesis. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Patel, M.; Kim, C.; Cheng, S.; Baral, C.; and Yang, Y. 2023. ECLIPSE: A Resource-Efficient Text-to-Image Prior for Image Generations. *arXiv preprint arXiv:2312.04655*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, 8821–8831. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2022. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494.

Saxon, M.; and Wang, W. Y. 2023. Multilingual Conceptual Coverage in Text-to-Image Models. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4831–4848. Toronto, Canada: Association for Computational Linguistics.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C. W.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; Schramowski, P.; Kundurthy, S. R.; Crowson, K.; Schmidt, L.; Kaczmarczyk, R.; and Jitsev, J. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Thrush, T.; Jiang, R.; Bartolo, M.; Singh, A.; Williams, A.; Kiela, D.; and Ross, C. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5238–5248.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

Xia, W.; Zhang, Y.; Yang, Y.; Xue, J.-H.; Zhou, B.; and Yang, M.-H. 2022. Gan inversion: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zarlenga, M. E.; Pietro, B.; Gabriele, C.; Giuseppe, M.; Giannini, F.; Diligenti, M.; Zohreh, S.; Frederic, P.; Melacci, S.; Adrian, W.; et al. 2022. Concept embedding models: Beyond the accuracy-explainability trade-off. In *Advances in Neural Information Processing Systems*, volume 35, 21400–21413. Curran Associates, Inc.

Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; and Metaxas, D. N. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 5907–5915.

Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; and Metaxas, D. N. 2018. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(8): 1947–1962.