GEOBIKED: A DATASET WITH GEOMETRIC FEATURES AND AUTOMATED LABELING TECHNIQUES TO ENABLE DEEP GENERATIVE MODELS IN ENGINEERING DESIGN.

Phillip Mueller BMW Group University of Augsburg phillip.mueller@bmw.de Sebastian Mueller BMW Group sebastian.mm.mueller@bmw.de

Lars Mikelsons University of Augsburg lars.mikelsonst@uni-a.de

ABSTRACT

Purpose – We provide a dataset for enabling Deep Generative Models (DGMs) in engineering design and propose methods to automate data labeling by utilizing large-scale foundation models.

Methodology – GeoBiked is curated to contain 4 355 bicycle images, annotated with structural and technical features and is used to investigate two automated labeling techniques:

- The utilization of consolidated latent features (Hyperfeatures) from image-generation models to detect geometric correspondences (e.g. the position of the wheel center) in structural images.
- The generation of diverse text descriptions for structural images. GPT-40, a vision-languagemodel (VLM), is instructed to analyze images and produce diverse descriptions aligned with the system-prompt.

Findings – By representing technical images as Diffusion-Hyperfeatures, drawing geometric correspondences between them is possible. The detection accuracy of geometric points in unseen samples is improved by presenting multiple annotated source images.

GPT-40 has sufficient capabilities to generate accurate descriptions of technical images. Grounding the generation only on images leads to diverse descriptions but causes hallucinations, while grounding it on categorical labels restricts the diversity. Using both as input balances creativity and accuracy.

Research implications – Successfully using Hyperfeatures for geometric correspondence suggests that this approach can be used for general point-detection and annotation tasks in technical images. Labeling such images with text descriptions using VLMs is possible, but dependent on the models detection capabilities, careful prompt-engineering and the selection of input information.

Originality – Applying foundation models in engineering design is largely unexplored. We aim to bridge this gap with a dataset to explore training, finetuning and conditioning DGMs in this field and suggesting approaches to bootstrap foundation models to process technical images.

Keywords Deep Generative Models, Data-driven design, AI-driven engineering design

Paper type Research paper

1 Introduction

Rapid advancements in the field of machine learning underscore the pivotal role of high-quality datasets in propelling technological breakthroughs. In Computer Vision, the introduction of quality, publicly available datasets has acted as a catalyst, enabling researchers to evaluate the performance of diverse methodologies. Datasets like ImageNet (Deng *et al.* 2009), CIFAR (Krizhevsky 2009) and MNIST (LeCun *et al.* 1998) are crucial to level the playing field and set benchmarks that define the state of the art.

Recent breakthroughs in generative tasks, such as the synthesis of high-quality natural images with Stable Diffusion (Rombach *et al.* 2022), rely on extensive, high-quality data pools such as LAION (Schuhmann *et al.* 2022) for model training. Methodologies for the conditional control of the synthesized content also require data. Training a single adapter to condition images on user-provided sketch-inputs within the ControlNet framework is based on 500k sketch-image-caption pairs (Zhang *et al.* 2023) obtained from an internet database. An amount that is often out of reach for limited, domain-specific applications of DGM's in engineering-design. Even approaches that are training-free require benchmark samples to evaluate the effectiveness of the mechanism.

Despite the public availability of general-purpose datasets for various deep learning applications, domain-specific fields like engineering design still face the scarcity of datasets equipped with detailed structural and geometric information. This gap has been highlighted in multiple studies (Regenwetter *et al.* 2022; Alam *et al.* 2024; Picard *et al.* 2023). It limits the application of Deep Generative Models (DGMs) in engineering design, where the objective extends beyond generating aesthetically pleasing visuals. Fulfillment of fundamental technical feasibility and unambiguous control over the process are prerequisites to develop real-world design concepts.(Alam *et al.* 2024; Joskowicz and Slomovitz 2024). In such scenarios, precise control over the generative process is necessary. Developing methods for conditioning and controlling the generation requires datasets with design-relevant and interpretable annotations.

In response to these challenges, our work ¹ builds upon and extends the foundational efforts of the BIKED project (Regenwetter *et al.* 2021). We streamline and enrich the dataset with geometric and semantic details aiming to enable engineers and designers to conduct basic experiments on DGMs with structural image data such as model training, finetuning, developing conditioning mechanisms and benchmarking. Our updated dataset includes interpretable features and geometric representations encompassing 12 reference points. We present the details about the dataset in Section 2.

Recognizing the bottleneck often encountered in annotating target data for generative tasks, our work furthermore explores the capabilities of utilizing off-the-shelf generative models with capabilities for visual- and languageunderstanding for automating these tedious manual efforts. We hypothesize that we can leverage the reasoning capabilities and context understanding learned by large-scale foundation models (Tang *et al.* 2023; Hwang *et al.* 2023; Picard *et al.* 2023; Singh *et al.* 2023) to bypass manual data annotation by passing this task to the generative model. In that, we investigate the capabilities of these models to process domain-specific, engineering design image data. Specifically, we ask two questions that relate to two relevant data annotation tasks.

- 1. Can we use the learned spatial and semantic understanding of pretrained latent diffusion models (Stable Diffusion (Rombach *et al.* 2022) to detect and annotate geometric correspondences in structural images?
- 2. Can state-of-the-art vision-language models (GPT-40 (OpenAI 2024b)) be used to generate diverse text descriptions of structural image data that accurately describe the technical object and its fine visual details?

The first question corresponds to the task of annotating the structural bicycle images with the geometric features, a task we did by hand for all samples. Analogous to how humans approach such tasks, we aim to provide only a handful of annotated examples. We then utilize the pretrained network by (Luo *et al.* 2023) to consolidate the learned feature representations, extracted from Stable Diffusion. These Hyperfeatures are used to draw correspondences between the geometric image features described by the annotated points and geometric features in unseen images. Leveraging the feature representations learned by Stable Diffusion, we automate the task of annotating geometric reference points in the GeoBiked images, showing that a) large-scale diffusion models for image generation can be used to process structural images in engineering design, b) a learned consolidation of latent features can be used to draw correspondences between such technical images, even if it has not been trained on them and c) presenting multiple reference images that show different styles of the design objects leads to higher accuracy in the prediction of the geometrical reference points in unseen images.

For the second question, we aim to generate diverse and creative text descriptions for the technical images by passing them to GPT-40 (OpenAI 2024b) as our off-the-shelf vision-language model (VLM). The VLM is instructed to produce descriptions of different lengths and styles. We investigate how the model can be aligned to produce desired descriptions of sufficient diversity without hallucination. Therefore, we evaluate the influence of three different configurations of input information. First, we only input the image for the VLM to describe. Second, we pass the labels annotating the corresponding image. The labels contain information about the style and the technical layout of each bicycle. Finally, we input both previous information types together.

To sum up our main contributions, we present the GeoBiked dataset that contains structural images of bicycles. The images are annotated with interpretable design, technical and geometric features relevant for engineering design applications. Based on the dataset, we propose two methods for automated labeling of structural image data utilizing off-the-shelf foundation models, that do not need any training or finetuning. First, we show that a learned consolidation

¹Dataset and Code is found under: https://anonymous.4open.science/r/GeoBIKED-D12C

of latent image features from Stable Diffusion can be used to accurately predict geometric reference points in unseen images. Second, with GPT-4o's vision-language capabilities, we generate diverse text descriptions of controllable style that accurately describe the structural images, balancing the needs for diversity and accuracy.

2 GeoBiked Dataset

2.1 Related Work

Our dataset is based on the BIKED-dataset (Regenwetter *et al.* 2021), which originated from the BikeCAD software (Curry 1998), a specialized tool for bicycle design. From an initial collection of 4791 bikes and 23813 descriptive parameters, the authors distilled 4512 bicycles and 1314 parameters. These parameters correlate to the CAD-models and describe every bicycle in detail. This curation process aimed to retain the raw richness of the data, facilitating a broad spectrum of data-driven design applications. Their methodology exemplifies the utility of the dataset through the training of two Variational Autoencoders (VAEs): one dedicated to generating new bike images and another to reconstructing bike parameters, showcasing the versatility of the dataset in supporting innovative design synthesis.

The initial version of the dataset comes with a number of shortcomings, limiting its applicability in DGM-driven engineering design tasks. Each sample is annotated with a total of 1314 parameters that were extracted from the BikeCAD software. Despite large in quantity, these parameters are largely non-interpretable and contain no meaningful information about design, style or structural composition of the bicycles they describe. Having such information is crucial to enable DGMs in engineering and concept design, as they are the basis for conditioning and control modalities of the generatice process (Alam *et al.* 2024; Joskowicz and Slomovitz 2024; Regenwetter *et al.* 2022). Furthermore, the dataset contains a number of infeasible designs and unrealistic samples. It lacks a uniform scaling of the objects within the image resolution, preventing direct geometric correspondences between the images.

2.2 Methodology

The first curation step in the derivation of GeoBiked from BIKED is a visual inspection of the 4512 provided bike images for faulty and out-of-distribution samples. We categorize samples as faulty if their geometric integrity is not ensured or the frame design is visibly unrealistic, see Figure Figure 1. We also remove the ten geometrically largest samples from the dataset as they are significantly out-of-distribution, therefore reducing the variance of the bike sizes and the geometric characteristics by more than 30% (Figure 8).



Figure 1. Unrealistic samples from the original BIKED dataset (Regenwetter et al. 2021)

The next step is the centering and geometric normalization of the images. In the original dataset, the images are not normalized to the same scale. Using the provided information about the scaling factor in the BIKED dataset, we ensure that all bikes are scaled equally. This step solidifies geometric consistency. Furthermore, we maximize the size of the bikes within the image resolution by precisely fitting the largest sample to the image resolution in length (x-dimension). This leads to one pixel equaling 10.19 mm in 256×256 images and 1.27 mm in 2048×2048 images.

To add geometric reference points to the dataset, we first define characteristic points and intersections in the bike geometry, aiming to allow the representation of each sample solely by a combination of these points. We describe the points in more detail in Section 2.3.3. Of the 12 geometric points, we derive six from the provided parameter set in the original dataset. The remaining parameters have to be defined manually. The coordinate-values of the geometric points are stored in millimeters for simplicity. They can be translated into pixel-values with the previously discussed scaling for both image resolutions. We filter and modify the initial set of descriptive parameters for features with semantic, geometric, or technical relevance, keeping a total of nine. We discuss the provided features in the next section in more detail.

2.3 Dataset Features

2.3.1 Design Features

We categorize the samples in the dataset into 19 different bicycle styles. This categorization is adapted from the original BIKED dataset. The style distribution is shown in Figure 2a. In the dataset, Road-Bikes are the most common style, followed by Mountain-Bikes and Track-Bikes.

We further add the diameters of the front and the rear wheel to the dataset. These are essential for the overall bicycle composition and geometry and can therefore be used as technical or design parameters. Furthermore, we provide the possibility of categorizing the samples by their Rim-style. The Rim-style has a significant impact on the overall design and appearance of the bicycle. The dataset distinguishes between spoked rims, tri-spoked rims and disked rims. Front-and rear wheel rim styles are separate categories, as several samples have a combination of two different rims. We note that spoked rims are by far the most common category in the dataset, making up 93.5% of samples for the front wheel and 92.9% for the rear wheel.

Another design-related feature that we provide in our dataset is the fork type. We distinguish between rigid forks, suspension forks and single-sided forks. Rigid forks make up 77.1% of samples while suspension forks make up 20.2% and single-sided forks 2.6%. The final set of design features are the bottles on the seat-tube and the down-tube. We added information about their presence as Boolean-values to the dataset.



Figure 2. Quantity distribution of bike styles, frame and tube sizes in GeoBiked.

2.3.2 Technical Features

To provide technical features, we categorize the samples by their tube-sizes, their frame-sizes and the number of teeth on the chainring. For the tube-sizes, each bicycle consists of four major tubes. To categorize the samples in a meaningful way, we calculate the average tube diameter for each of the four tubes. The average seat-tube-diameter is 31.5mm, the average down-tube-diameter is 35.5mm, the average head-tube is 42.9mm in diameter and the top-tube measures at 32.0mm on average. The classification is carried out by counting how many diameters of a given sample are greater than the average. If all diameters are smaller than their class-average ($n_{TS} = 0$), the sample is considered to have a "minimal" tube-size. For $n_{TS} = 1$, the tube-size is considered "lite" and "standard" for $n_{TS} = 2$. If $n_{TS} = 3$ the tubes are "reinforced" and for $n_{TS} = 4$ they are categorized as "extreme". The frequency of each category is visualized in Figure 2b.

In addition to the tube sizes, we provide information about the bicycles' frame sizes, as this is a common metric for categorization. The most common way to determine the frame size is to measure the length of the seat-tube between the bottom bracket and the top edge of the tube (Norman 2024). We consider all samples with a seat-tube length smaller than 360 mm to be "XS". For seat-tube lengths between 360mm and 420mm, the samples are considered "S" and for lengths between 420mm and 480mm they are considered size "M". A bicycle is of frame-size "L" if the seat-tube length is between 480mm and 540mm and "XL" if it is longer than 540mm. The distribution of frame sizes is shown in Figure 2c.

2.3.3 Geometric Features

With the selection of the 12 reference points, we aim to capture bicycles in all styles and sizes and characterize them by their geometrical layout. The final selection of 12 geometric points is shown in Figure Figure 3. We select the center points of the rear wheel (RWC) and the front wheel (FWC) for obvious reasons as they define the wheelbase of the bicycle. The point "BB" marks the center of the intersection of the seat-tube and the chain-stay and therefore the center of the bottom bracket. The head-tube-top (HTT) marks the upper end of the tube that connects the handlebars to the fork, given that this is where the stem intersects with the fork-tube. The stem-top (ST) marks the end of the

stem and is of significant influence for the reach of the bicycle frame (Norman 2024). We define another point on the front-fork (FF), characterizing its shape and potential bends or angles. The seat-tube-top (STT) marks the upper end of the seat-tube and the saddle-top (SAT) describes the highest point of the saddle.

In addition to the points already mentioned, we define every intersection of the tubes that make up the bicycle frame. Namely, the intersections of top-tube and seat-tube (TTST), top-tube and head tube (TTHT), down-tube and head tube (DTHT), rear-tube and seat-tube (RTST). All points are defined by their x- and y-coordinate values, given in millimeters. The values are relative to the rear-wheel-center, which we defined as the center of the local coordinate system. Since the position of RWC in the image varies from sample to sample, we provide two additional values, x_{zero} and y_{zero} , that describe the distance of RWC to the bottom-left corner of the image. This allows for accurate localization of the geometric points within an image as well as comparisons of the geometric layout and bicycle sizes.



Figure 3. Geometric layout described by the geometric reference points.

2.4 DGM-driven Engineering Design Applications

Our objective with GeoBiked is to enable DGMs for engineering design applications. We provide a dataset that can be used for initial experiments and investigations which then lead to a faster and more efficient realization of real-world engineering design use-cases. In this section, we briefly discuss potential applications of GeoBiked and give some examples.

Model verification. For the case of training or finetuning visual DGMs in engineering design applications, the dataset can be used to help find a suitable model architecture. By training different types of DGMs on GeoBiked, practitioners get valuable insights into the capabilities of DGMs to handle structural image data. Such experiments furthermore allow for the evaluation of required computational ressources and training durations as well as dataset quality and size. This lets the user draw quick correspondences about the suitability and the limitations of given models for the task at hand. As a comprehensible example of this application, we train three different visual DGMs on GeoBiked. We train a convolutional VAE (Kingma and Welling 2013; Sohn *et al.* 2015; Rombach *et al.* 2022), a variant of the same VAE with adversarial training (Blattmann *et al.* 2021), a simple diffusion model (DDIM) (Ho *et al.* 2020; Song *et al.* 2022; Fan *et al.* 2023a) and a latent diffusion model (LDM) (Rombach *et al.* 2022) to reproduce the structural images. A brief summary of the results is provided in Table I and Figure 4 shows some qualitative examples. Details about the implementation can be found in Section 6.

While both diffusion-based models are able to reproduce the structural image data, the VAE-architecture struggles to synthesize clean images. Even with the introduction of an adversarial loss, the results that can be synthesized remain very blurry. In terms of compute requirements, the VAE is much cheaper but its capabilities are not sufficient for detailed images with fine-grained structures. Even though there exist Autoencoder-based approaches for such image data (Fan *et al.* 2023b), we observe that off-the-shelf diffusion models handle this data much better. Comparing both diffusion-models, we observe that in both quantitative (FID-score (Szegedy *et al.* 2014)) and qualitative measures (Figure 4), the LDM is outperformed by the DDIM. The DDIM, although significantly more expensive in training, reproduces the fine-grained structures better. Due to the images being encoded into a latent representation before learning the diffusion model, the LDM loses some structural and geometric information and the results are more blurred.

Conditioning. For engineering design applications, the possibility to condition the DGM on relevant modalities is a fundamental requirement. Developing custom conditioning mechanisms either requires training data, for example to

Table I. Model feasibility study. We compare four different DGM-architectures to synthesize the GeoBiked image distribution. The training durations are normalized for 20GB VRAM. While VAE-architectures require significantly less compute, they do not allow for the generation of high quality outputs with structural details. The DDIM shows the best FID-score, but requires the most computational ressources.

Architecture	$FID\downarrow$	Normalized Training Duration (hr)
CVAE	123.85	0.23
Adversarial CVAE	166.98	0.25
DDIM	12.03	68.05
LDM	20.69	30.18



Figure 4. Generated samples with DGMs trained on GeoBiked.

train a ControlNet adapter (Zhang *et al.* 2023), or test data to optimize and verify the method. GeoBiked can help with these tasks. In a different work, the dataset is used to enable a training-free architecture for visualization of engineering design images in realistic scenes (Mueller *et al.* 2024).

Evaluation of pretrained models. The GeoBiked dataset also provides an opportunity to evaluate pretrained models with regards to their context understanding and capability to handle detailed, CAD-like image data. For example, users can use the dataset to verify if a pretrained vision-language model posses enough visual understanding and context knowledge to accurately categorize or describe the images. We discuss this task in detail in Section 3 and Section 4.

3 Geometric Feature Detection with Diffusion Hyperfeatures

Adding structural or geometric information to technical images is fundamentally important for the application of DGMs in engineering design. Despite the recent explosion in publications about visual DGMs, there still is a shortage of datasets with domain-specific modalities that enable conditional control over the generative process. However, this control is a key requirement for the successful application of DGMs for technical and design tasks (Alam *et al.* 2024; Mueller and Mikelsons 2024). In GeoBiked, we provide information about the geometric layout of each image sample. This might be used to investigate geometry-aware image generation or conditioning a model on a geometric layout. In annotating the geometric points largely by hand, we come to the conclusion that this time-intensive manual task poses a significant barrier for domain-specific applications. We therefore aim to utilize the spatial and semantic understanding inherited by pretrained, large-scale diffusion models for image generation (Stable Diffusion) and automate this task (Po *et al.* 2023; Tang *et al.* 2023; Luo *et al.* 2023).

3.1 Related Work

A number of works have recently addressed the idea of utilizing and transferring learned latent representations from large-scale models for downstream tasks (Goodwin *et al.* 2022; Oquab *et al.* 2024; Tumanyan *et al.* 2022). From various studies and applications, we know that those representations carry information about the underlying structure and composition of images (Fan *et al.* 2023a; Sabour *et al.* 2024). The possibility to draw semantic correspondences with these deep features has also been proven (Amir *et al.* 2022; Caron *et al.* 2021; Oquab *et al.* 2024).

Recently, Luo *et al.* 2023 have made significant progress in this domain by consolidating the latent feature maps from SD, extracted over multiple layers and timesteps, into an interpretable, per-pixel descriptor. The so-called Hyperfeature map allows to draw semantic correspondences between two images by comparing their respective Hyperfeature maps. In their work, they train an aggregation network to consolidate an image into its Hyperfeature Map. This is done by caching the intermediate feature maps obtained by either generating a synthetic image or inverting an existing image through multiple diffusion timesteps. Each extracted feature map is upsampled to a standard resolution and passed through a bottleneck layer for standardization. Subsequently, mixing weights are learned that identify the most significant features. For the task of semantic correspondence, the cosine similarity is computed between the flattened descriptor maps of image pairs that are labeled with corresponding keypoints. The aggregation network is trained by minimizing a symmetric cross-entropy loss similar to CLIP (Radford *et al.* 2021) between the predicted and ground truth keypoints.

3.2 Method

We propose to utilize the aggregation network pretrained for semantic correspondence to draw geometric correspondences in the GeoBiked images. In their work, Luo *et al.* 2023 already show that their approach outperforms existing alternatives for matching keypoints in natural images. We apply their methodology for our dataset and extend it to be able to handle multiple annotated source images. We draw inspiration from the manual labeling process, where a human annotator is shown a small number of annotated reference images that show the relevant features and subsequently identifies them in unseen images showing similar concepts. The possibility to base the prediction of the geometric reference points on multiple source images is introduced to make the prediction more reliable given the variety of bicycle geometries and styles in the dataset.

The source images $i \in I_s$ are annotated with their corresponding geometric points p_{ik} . It is worth noting that while the points can be chosen freely, they have to be consistent inbetween the source images, meaning that they mark the same geometric characteristics. Using the pretrained aggregation network, the Hyperfeature map $H_{s_i} \in \mathbb{R}^{C \times 64 \times 64}$ is computed for each source image. Depending on the resolution of the source images $(H \times W)$, the Hyperfeatures are interpolated back to the original image size to obtain $H'_{s_i} \in \mathbb{R}^{C \times H \times W}$. They are subsequently flattened and normalized to form $F_{s_i} \in \mathbb{R}^{(H \cdot W) \times C}$. Per source image, each labeled point $k \in p_{ik}$ is translated into its index idx_{ik} to extract the corresponding Hyperfeatures V_{s_i} from the flattened map F_{s_i} :

$$V_{s_i} = F_{s_i}[:, idx_{ik}, :] \text{ with } V_{s_i} \in \mathbb{R}^{N \times C}.$$

$$\tag{1}$$

When processing an entire dataset, the unlabeled images to be annotated are processed one after another. For each target image I_t , we compute its Hyperfeature map $H_t \in \mathbb{R}^{C \times 64 \times 64}$, interpolate it to the original image size $H'_t \in \mathbb{R}^{C \times H \times W}$ and also flatten and normalize it for $F_t \in \mathbb{R}^{(H \cdot W) \times C}$. Given the Hyperfeature representations of the source points V_{s_i} and the target image F_t , we now can compute the similarity matrix:

$$S_i = V_{s_i} \times F_t^T \text{ with } S_i \in \mathbb{R}^{N \times (H \cdot W)}.$$
(2)

We obtain one similarity matrix of size $N \times (H \cdot W)$ per combination of source and target image, where N is the number of points. The similarity matrix contains the cosine similarities between the Hyperfeatures of the annotated source points and the Hyperfeatures of the entire target image. Now we extract the maximum cosine similarities per row in S_i to obtain v_i , which describes the Hyperfeatures in the target image that have the highest correspondence to the Hyperfeatures of source image *i* with respect to the points. Since we are processing multiple source images, we can concatenate the similarity vectors \mathbf{v}_i along the y-axis and now extract the row-wise maximum similarity:

$$\mathbf{v}_{max}[k] = \max(\mathbf{v}_i[k]) \tag{3}$$

This increases flexibility in the prediction of the geometric points as per-point, the source-target combination with the highest correspondence is selected. In simple terms, this allows us to predict the positions of the geometric reference points in the target images using the information from all the source images. We always chose the position of the point where the correspondence between the Hyperfeatures of the source and target image is highest.

3.3 Experiments

In our experiments, we aim to verify the hypothesis that we can in fact use the diffusion Hyperfeatures to detect the geometric features in unseen structural images. Furthermore, we want to find the optimale selection and quantity of annotated source images to accurately label the diverse bicycle images in the dataset. For the evaluation, we select a subset of 150 diverse samples from our dataset. All calculations are conducted on an NVIDIA RTX A4500 with 20GB of VRAM. The results are summarized in Table II as well as in Figure 11. For evaluation, we use the pixel-wise MAE and MSE between the predicted location of the point and the ground truth location that we annotated by hand. We average the errors over all annotated points per image to gain insights on how well the entire geometric layout is captured and predicted.

Table II. Experiment results. The accuracy of the geometry-points detection in the GeoBiked subset depends on the selection and quantity of annotated source images. We evaluated multiple bicycle style combinations per quantity. The results below are the best results per source image quantity. The metrics are computed as a pixel-wise distance between prediction and ground truth and averaged over all 12 geometry-points. The duration measures the processing time of the subset of 150 samples.

Source Image Quantity	$MAE\downarrow$	$MSE\downarrow$	Duration (min.)
1	2.429	33.158	21.76
2	2.063	13.836	22.11
3	1.989	11.367	24.97
4	1.943	12.251	27.76
5	2.025	12.256	28.15
6	1.845	11.733	26.82
7	2.028	14.204	26.12
8	2.001	13.928	30.74

3.3.1 Results

Single Source Image. When providing only one source image, we observe that the accuracy of the point-location prediction is heavily dependant on the type of source image that is provided as reference (see Table IV and Figure 9). We selected various bicycle styles as reference images and observed that an average geometry, which captures a wide selection of bicycle frames, leads to good prediction accuracy as the MAE is below 3 pixels. However, with this kind of source image, the method is not able to draw accurate correspondences to bicycles with a significantly different geometric layout. This is evident in the large MSE. When samples are chosen as source images that are not in the middle of the geometric distribution, the prediction accuracy deteriorates significantly, as seen in Table IV. For the source image with the best prediction accuracy, we still observe typical error patterns. Tube intersections are often annotated inaccurately as well as the saddle top and points around the handlebars. For the annotation of outlier samples, only the wheel centers are captured reliably.

The observed ambiguities most likely originate from the structure of the bicycles in the images. Due to them being plain grayscale structures on a white background, areas with tube intersections look very similar. A single source image does not capture the different options of the bicycle layout with respect to the saddle position, stem and handlebars and tube intersections. For example, if multiple tube intersections fall into the same position in the source image, but are in different positions in the target image, they will not be detected with high precision.

Multiple Source Images. Comparing an unlabeled target image with multiple source images for geometric correspondence noticeably improves the accuracy of the point prediction. By just using a second source image showing a different type of bicycle, we reduce the MSE by a factor of 2.4. The variety of layouts in the dataset gets captured much more reliable. Since we can precompute the Hyperfeatures for the source images once and then use them for processing the entire dataset, the computational overhead is insignificant. For two source images, we test a variety of combinations. We observe that using the sample that performs best when used as the only input together with a sample showing a different, but common, style leads to the best accuracy. Typical error patterns observed in the previous section are largely eliminated (see Figure 5 and Figure 10). Nevertheless, for uncommon styles the prediction is still inaccurate.

Based on the observations with one and two source images, we add a third source image showing such sample. With that, we are able to further reduce the MSE by about 18%. In our experiments, we observe that increasing the number of source images beyond three does not have a significant impact on the prediction accuracy (see Table II and Figure 11). We therefore propose that, in the case of the GeoBiked dataset, using three source images presents a good balance of

manual annotation effort and prediction accuracy of the automated process. Using the three source images shown in Figure 5, we also process the entire GeoBiked dataset, achieving an accuracy of MAE = 1.837 and MSE = 14.009 in a processing duration of 11.33 hours.



Figure 5. Qualitative Comparison of error patterns using one, two and three source images respectively. Images in the left columns (red marks) are source images. Images in the right column (blue marks) are target images. Most of the inaccuracies disappear when using three source images compared to a single one. Some uncertainty remains when uncommon samples are processed (see middle and last column). Best viewed when zoomed in.

3.3.2 Implications

For engineering design images with fine-grained details, the provided source images have to capture the variety of structures in the target image space as much as possible. It is important to select the references so that they show different layouts of the object geometry. In our case, this correlates to bicycles with overlapping and non-overlapping tube intersections. When processing a diverse dataset, outliers and uncommon samples require specific attention. They need to be annotated by and or at least evaluated for sufficient accuracy. When these limitations are regarded for, pretrained diffusion models can be used to automate the process of data annotation. Finetuning of the aggregation network might be necessary for some applications where the off-the-shelf accuracy is not sufficient.

Our experiments show that representing images through their diffusion Hyperfeatures generally allows us to process domain-specific, structural image data. Even though the diffusion model was not specifically trained on such data, the learned semantic correspondences and spatial understanding can be transferred and applied to engineering design tasks. Efficiently using the inherited capabilities however is not straightforward, as the feature representations extracted from the diffusion model need to be further processed. In this case, training an encoder-like model is required to draw semantic correspondences. Nevertheless, compared to a full-scale training or finetuning of a diffusion model this is relatively cheap. The aggregation network was trained on a single NVIDIA RTX TITAN 24GB GPU (Luo *et al.* 2023). Computing the Hyperfeatures of an image and using them for downstream applications can also be considered efficient as it is possible on a consumer-grade GPU.

In an outlook, we propose that Diffusion Hyperfeatures of images can also be used for other engineering design applications in image generation and modification. A future research direction can be to utilize them in order to improve object consistency in generative processes by comparing the Hyperfeature representations of the generated object with a learned distribution or ground-truth example. In a different work building up upon Diffusion Hyperfeatures, the authors address this task in broad terms (Luo *et al.* 2024).

4 Automatic Generation of Text Descriptions with Vision-Language Models

The possibility of text-based conditioning is a key factor of the recent success in image generation as it provides an intuitive modality to control the generated content. Numerous works discuss this topic. Most notably (Ramesh *et al.* 2021; Rombach *et al.* 2022; Saharia *et al.* 2022b; Ruiz *et al.* 2022; Inc. 2023) in image generation and (Saharia *et al.* 2022a; Brooks *et al.* 2023; Yang *et al.* 2023; Tumanyan *et al.* 2022) in image-to-image editing. It provides a way to

dynamically adjust the amount of information passed to the model. For engineering design applications, the DGM can be trained to adhere to many design constraints and requirements as well as fill in the blanks for short, high-level descriptions. Training such models and corresponding conditioning mechanisms requires pairs of images and diverse text-descriptions. For better generalisation, the text-descriptions have to contain varying amounts of information.

Labeling domain-specific images with diverse text-prompts by hand is generally unfeasible. In addition to the high manual effort on a repetitive task, the descriptions would follow the bias of the annotators and be limited in their diversity and creativity. Therefore, we provide a recipe on how to bootstrap the capabilities of large-scale (vision-) language models to automate this task. We employ GPT-40 (OpenAI 2024b), which build upon GPT-4 (OpenAI 2024a) and GPT-4(Vision) (OpenAI 2023) with the additional benefit of providing noticeably faster inference. We use this model because of its state-of-the-art vision-language understanding and reasoning and its easy accessibility through the API. In addition to providing a method for automatic labeling, this task provides insights on the capabilities of GPT-40 for context understanding and reasoning in technical images with fine-grained details.

4.1 Related Work

Our approach fundamentally is an image captioning task. Previous works address this topic with a variety of approaches, utilizing CLIP-embeddings (Mokady *et al.* 2021) or vision-language transformers (Zhou *et al.* 2019; L. H. Li *et al.* 2019). The development of GPT-4(V) is a major improvement as it possesses vast context understanding and broad general reasoning capabilities to accurately caption contents of an image (Singh *et al.* 2023; Hwang *et al.* 2023). Besides the multi-modal variations of GPT-4, other vision-language models like LLaVA (Liu *et al.* 2023; Liu *et al.* 2024) are available, but are outperformed on the VisIT-Benchmark for VLMs (Bitton *et al.* 2023) by GPT-4(V) in terms of model performance across a diverse set of instruction-following tasks.

The comprehensive study by Picard *et al.* 2023 investigates the feasibility of including GPT-4(V) in an engineering design process and thereby underlines our assumptions that the capabilities of the model are sufficient for domain-specific captioning tasks on technical images.

In terms of using language models for generating synthetic descriptions of images, a few works are to be named. Cosmopedia (Ben Allal *et al.* 2024) is a dataset of synthetic textbooks, blogposts, stories, posts and WikiHow articles generated by Mixtral-8x7B-Instruct (Jiang *et al.* 2023). The dataset contains over 30 million files and 25 billion tokens.

BLIP-2 employs a two-stage pre-training approach to enhance image-grounded text generation. By using transformers for both image and text processing, BLIP-2 generates accurate and contextually relevant descriptions from images. It demonstrates the ability to perform zero-shot image captioning effectively, making it versatile for various applications (J. Li *et al.* 2023).

LAVIS is a whole suite for developing multi-modal models, allowing to rapidly employ and benchmark models for different tasks, image captioning being one of them (D. Li *et al.* 2023). DreamSync (Sun *et al.* 2024) employs VLMs to automate the selection of high-quality image-text pairs to finetune a text-to-image model. Garg *et al.* 2024 employ a VLM together with human annotators to create a dataset of hyper-detailed, synthetic image descriptions.

4.2 Method

We access GPT-40 through the API and prompt it to generate descriptions of the GeoBiked bicycle images. To obtain diverse descriptions of the images, we prompt GPT-40 to adhere to specific description characteristics. The system-prompt passed to GPT-40 specifically outlines the format of the generated description. The mask for the system-prompt is provided in Appendix C.1.

- Length: We distinguish between short (5-10 words), medium (10-20 words) and long (20-40 words) descriptions.
- Vibe: The model is prompted to generate descriptions with either a technical character or a casual description.
- **Style:** The descriptions have to be either in style of a marketing message or of a prompt for a text-to-image model (Midjourney).

We generate text-descriptions with three different data sources that are provided to GPT-40. For the first one, only the bicycle image is provided. It is wrapped in the system-prompt after the general task description and the definition of the required description characteristics. The general task instructs to model to analyse the input information and create a description for it while adhering to the formulated description characteristics. The second data source are labels in the form of technical categories describing the bike. No image is used for this configuration. From GeoBiked, we use bicycle style, rim styles, fork type and whether there are bottles on the down- or seattubes. In the third configuration, we pass the image together with the technical categories.

The motivation for selecting the three input information combinations is grounded on several ideas. By providing only the image, we generally test if the VLM has sufficient context understanding and visual detection capabilities to capture the details of the presented bicycle images. The label-only configuration is used for comparison and for verification that the increased information density of an image compared to five categorical features actually leads to more diverse and accurate descriptions. With the third setup, we test if both information inputs used together hold an advantage over any combination on its own.

4.3 Experiments

First and foremost, we aim to verify if GPT-40 is able to generate creative and diverse descriptions of the samples from GeoBiked, given that the images are of a technical character and contain fine-grained visual details. We further analyze if the generated descriptions follow the required length, vibe and style. By also passing ground-truth labels to the model, we evaluate the possibility of generating diverse, but hallucination-free descriptions as hallucination is a well-known issue with LLMs (Huang *et al.* 2023).

In terms of evaluation metrics, we measure diversity by counting the number of unique outputs generated with different description requirements. This metric provides insights on the repetitiveness of the generated descriptions and therefore indirectly measures how well the VLM can capture different inputs and create unique descriptions from them. Additionally, we compute the Levenshtein-distance between unique outputs as a measurement for the difference between two linguistic sequences (Haldar and Mukhopadhyay 2011). A higher average Levenshtein-distance indicates more diverse text-descriptions. We use the implementation by (Bachmann 2024).

To evaluate the accuracy of the generated text-description with the ground-truth labels, we again use GPT-40. In this setting, the model is utilized as a classifier. We instruct it to infer the categorical labels from the text description and then compare the extracted labels from the text-description to the ground-truth labels and count the inconsistencies. The code is provided in the linked repository. We acknowledge that this approach adds a degree of uncertainty due to the classifier having to infer the labels from the text-description. We assume that the error injected by this process is consistent and we therefore can still derive correlations between the system-prompt configuration, the provided input information and the accuracy of the generated text-prompt.

We conduct our experiments on the entire GeoBiked dataset, containing all 4355 samples. We use the images with 2048×2048 resolution. For better comparison, we only use one thread, calling the API one time after another. In terms of hardware, for CPU we use 12th Gen Intel Core i7-12850HX 2.10 GHz. Generating the descriptions for each of the three input configurations takes approximately 15 hours.

4.3.1 Results

Our results for description uniqueness are visualized in Figure 6. When only the image is passed to GPT-40, we observe that for short descriptions, between 84.5% and 88.2% of the outputs are unique. Short descriptions therefore tend to be repetitive in some cases (see Figure 6a). The long descriptions are almost entirely unique (between 96.9% and 99.9%). On average, we obtain 90.9% unique descriptions. This is also the case when an image together with the ground-truth label is passed (Figure 6c). When the ground-truth labels are used as sole input, the uniqueness of the generated descriptions is significantly reduced for the short ones (Figure 6b). Even if the temperature, a parameter for randomness and therefore diversity of outputs, is set to a high value of 1, the label grounding restricts the model so much that the uniqueness is noticeably reduced.



Figure 6. Percentage of unique descriptions generated by GPT-40, compared over different configurations of description characteristics. Best viewed zoomed in.

The analysis of the Levenshtein distance as a measure of diversity and creativity supports the previous findings. Providing only an image leads to a greater Levenshtein distance among unique text-descriptions, as shown in the violin plot (Figure 7). For image-grounded text-descriptions, the average Levenshtein distance is significantly higher than for label-grounded descriptions, especially for long, technical ones. Grounding the generation on image and labels seems to also restrict the diversity of the unique outputs. As this configuration of input information generally produces more unique outputs than label-only grounding, we assume that providing the image leads to diverse descriptions while the label streamlines the descriptions in terms of format and sentence structure, hence the lower Levenshtein distance between unique descriptions.

For short descriptions, the distances of the unique outputs are again greater for image-only grounding, although the difference is smaller. A fact to keep in mind here is that image-grounding produces far more unique outputs in total.



Figure 7. Levenshtein distances between unique values for all three configurations of inputs.

When manually evaluating the generated text descriptions, we find that the ones generated on image grounding are more creative overall. For the label-grounded descriptions, we observe a pattern that they tend to be a concatenation of the ground-truth labels wrapped with some filling text. Furthermore, some design characteristics of bicycles are not captured by the ground-truth labels and therefore not regarded for in the generated description. Examples of this can be found in Table V. The bicycle in column four is correctly identified and described as vintage or retro when using the image for grounding, while the label-grounded descriptions do not mention anything about the extraordinary design. The same is true for the example given in column five, where the pannier, that is clearly visible in the image, is correctly mentioned for both image-grounded descriptions, but not for the label-based descriptions.

Besides diversity in generated outputs, we measure the accuracy of the generated descriptions to verify that they fit the image they aim to describe. The accuracy is calculated by instructing GPT-40 as a classifier and count the quantity of features that are described inaccurately, as described in Section 4.3. For the results in Table III, the error quantities are averaged over all text-descriptions of an instruction-setting (e.g. Length: Long, Vibe: Casual, Style: Marketing-Message). The results underline our previous findings. Providing only an image as input information leaves more degrees of freedom for the description generation. While this leads to more diverse outputs, it also increases hallucination. Shorter descriptions seem to be more accurate than longer ones for image-only grounding. We observe a mean error of 1.048 for short descriptions and 1.215 for long descriptions, as the model has more room for hallucinations. On average, there are 1.13 labels described inaccurately per description.

Only providing the ground-truth labels restricts the diversity of outputs but therefore noticeably limits hallucinations. The average error per description is reduced by a factor of 5.38. If an image is passed together with the ground-truth label, the accuracy is lowered by about 28.6% compared to label-only, but still surpasses the image-only accuracy by a factor of 4. The results are summarized in Table III. For these two cases, we observe that longer prompts are more accurate than shorter ones. For label-only inputs, short descriptions contain 0.265 errors on average while long descriptions contain 0.15. For image and labels as the data source, short descriptions lead to a mean error of 0.355 per description and long ones to only 0.1725.

Table III. Average error ratios for generated text-descriptions compared to the ground-truth annotations in GeoBiked. We compare different configurations of the system-prompt (Instruction Setting). The table shows the average number of categorical labels that the text description misrepresented per instruction setting and input information configuration.

Instruction Setting	Image Only	Ground-Truth Label	Image and Label
Long; Casual; Marketing-message	0.72	0.19	0.20
Long; Casual; Prompt-to-midjourney	1.05	0.20	0.27
Long; Technical; Marketing-message	1.33	0.13	0.12
Long; Technical; Prompt-to-midjourney	1.76	0.08	0.10
Short; Casual; Marketing-message	0.77	0.36	0.40
Short; Casual; Prompt-to-midjourney	1.03	0.31	0.34
Short; Technical; Marketing-message	1.08	0.23	0.30
Short; Technical; Prompt-to-midjourney	1.31	0.16	0.38
Mean	1.13	0.21	0.27

4.3.2 Implications

For the task of automatically annotating technical images with feasible and diverse text descriptions, GPT-40 is an adequate choice. Due to the broad context understanding and the sufficient vision-language reasoning capabilities, domain-specific details are captured and described accurately. For the configuration of the input information, there exists a tradeoff between diversity of the generated text descriptions and their accuracy compared to ground-truth labels. While grounding the generation on labels leads to more accurate descriptions and reduces hallucinations, it restricts the diversity. For our example, pure label-grounding also leads to loss of information, as the labels contain significantly less details than the images. Providing an image paired with labels appears to reduce hallucinations while accurately capturing most visual features of the geometries. The diversity of text descriptions is noticeably reduced compared to only using an image as input information. Therefore, this configuration is somewhat of a tradeoff between diversity and accuracy. The configuration of instruction settings and input information ultimately depends on the desired application of the generated text descriptions.

Although GPT-4o reliably adheres to the instructions stated in the system-prompt, we find that for large-scale automated annotation tasks, the system-prompt has to be evaluated and systematically optimized. The requirements and constraints for the model need to be explicitly included to avoid hallucinations. The system-prompt severely influences the quality of the outputs and their downstream usability. The same is true for employing GPT-4o as a label-classifier to infer categories from natural language texts. The model possesses sophisticated capabilities for this task that can be of benefit, as no use-case specific classifier has to be trained on the given problem. However, this does require careful prompt-engineering for the outputs to be in a structure that can be evaluated algorithmically. We find the DSPy-library of enormous help for such tasks (Khattab *et al.* 2023).

To process technical or CAD-like images with fine-grained visual details and generate adequate descriptions, large VLMs are necessary. In an ablation experiment, we instruct the significantly smaller Moondream model (vikhyatk 2024) with the same task and observe that the generated descriptions are very uniform. Unique descriptions are generated rarely. The descriptions that are unique show a low Levenshtein distance, as all of them follow the same sentence structure (see Figure 12).

5 Limitations and Future Work

GeoBiked Dataset. With the dataset, we aim to build a foundation to apply DGMs in engineering design processes. While real-world use-cases require data beyond bicycle images, comparing the performance of DGMs is an important first-step to foster the future development of the field. GeoBiked presents an opportunity to create benchmark-challenges for visual DGMs. There is however a need for evaluation metrics that capture the structural integrity and real-world feasibility of the synthesized images. While there are approaches to better capture structural feasibility in images (Fan *et al.* 2024; Regenwetter *et al.* 2023), this largely remains an open challenge. Additionally, the dataset currently does not include any explicit features to evaluate the performance of the bicycle designs. Adding this is another step towards improving applicability in engineering design domains.

Geometric Feature Detection with Diffusion Hyperfeatures. While we show that Diffusion Hyperfeatures can be used to draw semantic correspondences in technical images and label the data in a few-shot fashion, the technique most likely requires refinement for more complex structures. Possible directions are retraining or finetuning the aggregation network for use-case specific problems or modifying the similarity computation to compare Hyperfeatures of patches rather than singular points. Overall, we find that is method is an exiting approach worth exploring for other problems such as object consistency and similarity in image and 3D generation.

Automatic Generation of Text Descriptions with VLMs. The dependence on GPT-40, as a model only accessible through the API without open-source weights is a drawback. While this solution is easy-to-use, we encourage investigating open-source alternatives such as LLaVA (Liu *et al.* 2023; Liu *et al.* 2024) for this task. Vision-language understanding capabilities are likely to improve rapidly in open-source alternatives.

6 Conclusion

We illustrate the potential of utilizing large-scale pretrained generative models to address the challenges of domainspecific tasks in engineering design. The GeoBiked dataset serves as a resource for baseline investigations into model feasibility, data requirements, and benchmarking in engineering design applications, providing a foundation for exploration of Deep Generative Models (DGMs). Additionally, we have demonstrated that off-the-shelf visual foundation models, when used effectively and guided correctly, can automate data labeling and annotation tasks, significantly lowering the barriers to entry for DGMs in technical fields.

Diffusion models for image generation inherit spatial and semantic understanding that can be used to draw geometric correspondences in structural images by consolidating image features into an interpretable Hyperfeature map. The prediction of geometric reference points in unseen images of technical objects is significantly improved if multiple examples, which are showing different styles of objects are used as reference. Large vision-language models are applicable to automatically generate text descriptions of technical images. The accuracy and creativity of the generated descriptions depend on careful prompting and the provided input information. Overall, our findings aim to facilitate the broader adoption of AI-driven approaches in engineering design, streamlining processes and expanding creative possibilities in the field.

References

- Alam, Md Ferdous, Lentsch, Austin, Yu, Nomi, Barmack, Sylvia, Kim, Suhin, Acemoglu, Daron, Hart, John, Johnson, Simon, and Ahmed, Faez (Mar. 2024). From Automation to Augmentation: Redefining Engineering Design and Manufacturing in the Age of NextGen-AI. In: *An MIT Exploration of Generative AI*. DOI: 10.21428/e4baedd9. e39b392d. (Visited on 04/02/2024).
- Amir, Shir, Gandelsman, Yossi, Bagon, Shai, and Dekel, Tali (2022). Deep vit features as dense visual descriptors. In: *Proceedings of the European Conference on Computer Vision Workshops (ECCVW)*.

Bachmann, Max (2024). Levenshtein.

- Ben Allal, Loubna, Lozhkov, Anton, Penedo, Guilherme, Wolf, Thomas, and Werra, Leandro von (2024). *Cosmopedia*. URL: https://huggingface.co/datasets/HuggingFaceTB/cosmopedia.
- Bitton, Yonatan, Bansal, Hritik, Hessel, Jack, Shao, Rulin, Zhu, Wanrong, Awadalla, Anas, Gardner, Josh, Taori, Rohan, and Schmidt, Ludwig (2023). VisIT-Bench: A Benchmark for Vision-Language Instruction Following Inspired by Real-World Use. arXiv: 2308.06595 [cs.CL].
- Blattmann, Andreas, Milbich, Timo, Dorkenwald, Michael, and Ommer, Björn (2021). iPOKE: Poking a Still Image for Controlled Stochastic Video Synthesis. arXiv: 2107.02790v2. (Visited on 07/15/2024).
- Brooks, Tim, Holynski, Aleksander, and Efros, Alexei A. (2023). InstructPix2Pix: Learning to Follow Image Editing Instructions. In: *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Caron, Mathilde, Touvron, Hugo, Misra, Ishan, Jégou, Hervé, Mairal, Julien, Bojanowski, Piotr, and Joulin, Armand (2021). Emerging properties in self-supervised vision transformers. In: *Proceedings of the International Conference on Computer Vision (ICCV)*.

Curry, Brent (1998). BikeCAD. URL: https://www.bikecad.ca/ (visited on 07/10/2024).

Deng, Jia, Dong, Wei, Socher, Richard, Li, Li-Jia, Kai Li, and Li Fei-Fei (June 2009). ImageNet: A Large-Scale Hierarchical Image Database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL: IEEE, pp. 248–255. ISBN: 978-1-4244-3992-8. DOI: 10.1109/CVPR.2009.5206848. (Visited on 12/15/2023).

- Fan, Jiajie, Trigui, Amal, Bäck, Thomas, and Wang, Hao (2024). Enhancing Plausibility Evaluation for Generated Designs with Denoising Autoencoder. arXiv: 2403.05352 [cs.CV]. URL: https://arxiv.org/abs/2403.05352.
- Fan, Jiajie, Vuaille, Laure, Bäck, Thomas, and Wang, Hao (Nov. 2023a). On the Noise Scheduling for Generating Plausible Designs with Diffusion Models. arXiv: 2311.11207 [cs]. (Visited on 04/12/2024).
- Fan, Jiajie, Vuaille, Laure, Wang, Hao, and Bäck, Thomas (July 2023b). Adversarial Latent Autoencoder with Self-Attention for Structural Image Synthesis. arXiv: 2307.10166 [cs, eess]. (Visited on 12/15/2023).
- Garg, Roopal, Burns, Andrea, Ayan, Burcu Karagol, Bitton, Yonatan, Montgomery, Ceslee, Onoe, Yasumasa, Bunner, Andrew, Krishna, Ranjay, Baldridge, Jason, and Soricut, Radu (2024). ImageInWords: Unlocking Hyper-Detailed Image Descriptions. arXiv: 2405.02793 [cs.CV].
- Goodwin, Walter, Vaze, Sagar, Havoutis, Ioannis, and Posner, Ingmar (2022). Zero-shot category-level object pose estimation. In: *Proceedings of the European Conference on Computer Vision (ECCV)*,
- Haldar, Rishin and Mukhopadhyay, Debajyoti (2011). Levenshtein Distance Technique in Dictionary Lookup Methods: An Improved Approach. In: ArXiv abs/1101.1232. URL: https://api.semanticscholar.org/CorpusID: 1049293.
- Ho, Jonathan, Jain, Ajay, and Abbeel, Pieter (Dec. 2020). Denoising Diffusion Probabilistic Models. In: *Neural Information Processing Systems*. arXiv. arXiv: 2006.11239 [cs, stat]. (Visited on 12/15/2023).
- Huang, Lei, Yu, Weijiang, Ma, Weitao, Zhong, Weihong, Feng, Zhangyin, Wang, Haotian, Chen, Qianglong, Peng, Weihua, Feng, Xiaocheng, Qin, Bing, and Liu, Ting (2023). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. In: ArXiv abs/2311.05232. URL: https://api. semanticscholar.org/CorpusID:265067168.
- Hwang, Alyssa, Head, Andrew, and Callison-Burch, Chris (Nov. 2023). Grounded Intuition of GPT-Vision's Abilities with Scientific Images. arXiv: 2311.02069 [cs]. (Visited on 02/06/2024).
- Inc., Midjourney (2023). Midjourney.
- Jiang, Albert Q., Sablayrolles, Alexandre, Mensch, Arthur, Bamford, Chris, Chaplot, Devendra Singh, Las Casas, Diego de, Bressand, Florian, Lengyel, Gianna, Lample, Guillaume, Saulnier, Lucile, Lavaud, Lélio Renard, Lachaux, Marie-Anne, Stock, Pierre, Scao, Teven Le, Lavril, Thibaut, Wang, Thomas, Lacroix, Timothée, and Sayed, William El (2023). Mistral 7B. In: abs/2310.06825. DOI: 10.48550/ARXIV.2310.06825. arXiv: 2310.06825. URL: https://doi.org/10.48550/arXiv.2310.06825.
- Joskowicz, Jose and Slomovitz, Daniel (2024). Engineers' Perspectives on the Use of Generative Artificial Intelligence Tools in the Workplace. In: *IEEE Engineering Management Review* 52.1, 258–267. DOI: 10.1109/EMR.2023. 3333794.
- Khattab, Omar, Singhvi, Arnav, Maheshwari, Paridhi, Zhang, Zhiyuan, Santhanam, Keshav, Vardhamanan, Sri, Haq, Saiful, Sharma, Ashutosh, Joshi, Thomas T., Moazam, Hanna, Miller, Heather, Zaharia, Matei, and Potts, Christopher (2023). DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines. In: *arXiv:2310.03714*.
- Kingma, Diederik P. and Welling, Max (2013). Auto-Encoding Variational Bayes. In: *ICLR 2014*. arXiv. eprint: 1312.6114 (cs, stat). (Visited on 12/15/2023).
- Krizhevsky, Alex (2009). Learning Multiple Layers of Features from Tiny Images.
- LeCun, Yann, Bottou, L, Bengio, Yoshua, and Haffner, P (1998). Gradient-Based Learning Applied to Document Recognition. In: *Proceedings of the IEEE* 86.11, 2278–2324.
- Li, Dongxu, Li, Junnan, Le, Hung, Wang, Guangsen, Savarese, Silvio, and Hoi, Steven C.H. (July 2023). LAVIS: A One-stop Library for Language-Vision Intelligence. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*. Toronto, Canada: Association for Computational Linguistics, pp. 31–41. URL: https://aclanthology.org/2023.acl-demo.3.
- Li, Junnan, Li, Dongxu, Savarese, Silvio, and Hoi, Steven (2023). BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. arXiv: 2301.12597v3 [cs.CV].
- Li, Liunian Harold, Yatskar, Mark, Yin, Da, Hsieh, Cho-Jui, and Chang, Kai-Wei (Aug. 2019). VisualBERT: A Simple and Performant Baseline for Vision and Language. arXiv: 1908.03557 [cs]. (Visited on 02/15/2024).
- Liu, Haotian, Li, Chunyuan, Li, Yuheng, and Lee, Yong Jae (May 2024). Improved Baselines with Visual Instruction Tuning. arXiv: 2310.03744 [cs]. (Visited on 07/09/2024).
- Liu, Haotian, Li, Chunyuan, Wu, Qingyang, and Lee, Yong Jae (Dec. 2023). Visual Instruction Tuning. arXiv: 2304. 08485 [cs]. (Visited on 07/09/2024).

- Luo, Grace, Darrell, Trevor, Wang, Oliver, Goldman, Dan B, and Holynski, Aleksander (2024). Readout Guidance: Learning Control from Diffusion Features. In: *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Luo, Grace, Dunlap, Lisa, Park, Dong Huk, Holynski, Aleksander, and Darrell, Trevor (May 2023). Diffusion Hyperfeatures: Searching Through Time and Space for Semantic Correspondence. arXiv: 2305.14334 [cs]. (Visited on 12/15/2023).
- Mokady, Ron, Hertz, Amir, and Bermano, Amit H (2021). ClipCap: CLIP Prefix for Image Captioning. In: *arXiv* preprint arXiv:2111.09734.
- Mueller, Phillip and Mikelsons, Lars (2024). Exploring the Potentials and Challenges of Deep Generative Models in Product Design Conception. arXiv: 2407.11104. (Visited on 07/15/2024).
- Mueller, Phillip, Wiese, Jannik, Craciun, Ioan, and Mikelsons, Lars (2024). InsertDiffusion: Identity Preserving Visualization of Objects through a Training-Free Diffusion Architecture. arXiv: 2407.10592. (Visited on 07/15/2024).
- Norman, Paul (2024). How to Measure a Bike Frame: Our Complete Guide to Sizing a Bike.
- OpenAI (2023). GPT-4V(Ision) System Card. Tech. rep.
- (2024a). GPT-4 Technical Report. arXiv: 2303.08774.
- (2024b). GPT-4o. URL: https://openai.com/index/hello-gpt-4o/ (visited on 08/08/2024).
- Oquab, Maxime, Darcet, Timothée, Moutakanni, Théo, Vo, Huy, Szafraniec, Marc, Khalidov, Vasil, Fernandez, Pierre, Haziza, Daniel, Massa, Francisco, El-Nouby, Alaaeldin, Assran, Mahmoud, Ballas, Nicolas, Galuba, Wojciech, Howes, Russell, Huang, Po-Yao, Li, Shang-Wen, Misra, Ishan, Rabbat, Michael, Sharma, Vasu, Synnaeve, Gabriel, Xu, Hu, Jegou, Hervé, Mairal, Julien, Labatut, Patrick, Joulin, Armand, and Bojanowski, Piotr (Feb. 2024). DINOv2: Learning Robust Visual Features without Supervision. arXiv: 2304.07193 [cs]. (Visited on 06/05/2024).
- Picard, Cyril, Edwards, Kristen M., Doris, Anna C., Man, Brandon, Giannone, Giorgio, Alam, Md Ferdous, and Ahmed, Faez (Nov. 2023). From Concept to Manufacturing: Evaluating Vision-Language Models for Engineering Design. arXiv: 2311.12668 [cs]. (Visited on 04/02/2024).
- Po, Ryan, Yifan, Wang, Golyanik, Vladislav, Aberman, Kfir, Barron, Jonathan T., Bermano, Amit H., Chan, Eric Ryan, Dekel, Tali, Holynski, Aleksander, Kanazawa, Angjoo, Liu, C. Karen, Liu, Lingjie, Mildenhall, Ben, Nießner, Matthias, Ommer, Björn, Theobalt, Christian, Wonka, Peter, and Wetzstein, Gordon (Oct. 2023). State of the Art on Diffusion Models for Visual Computing. arXiv: 2310.07204 [cs]. (Visited on 04/29/2024).
- Radford, Alec, Kim, Jong Wook, Hallacy, Chris, Ramesh, Aditya, Goh, Gabriel, Agarwal, Sandhini, Sastry, Girish, Askell, Amanda, Mishkin, Pamela, Clark, Jack, Krueger, Gretchen, and Sutskever, Ilya (2021). Learning Transferable Visual Models From Natural Language Supervision. In: *Proceedings of the 38th International Conference on Machine Learning (PMLR)*. arXiv:2103.00020. arXiv. arXiv: 2103.00020 [cs]. (Visited on 12/15/2023).
- Ramesh, Aditya, Pavlov, Mikhail, Goh, Gabriel, Gray, Scott, Voss, Chelsea, Radford, Alec, Chen, Mark, and Sutskever, Ilya (18–24 Jul 2021). Zero-Shot Text-to-Image Generation. In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Meila, Marina and Zhang, Tong. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 8821–8831. URL: https://proceedings.mlr.press/v139/ramesh21a.html.
- Regenwetter, Lyle, Curry, Brent, and Ahmed, Faez (2021). BIKED: A Dataset and Machine Learning Benchmarks for Data-Driven Bicycle Design. In: *International Design Engineering Technical Conference & Computers and Information in Engineering*. DOI: 10.1115/DETC2021-71681.
- Regenwetter, Lyle, Nobari, Amin Heyrani, and Ahmed, Faez (Mar. 2022). Deep Generative Models in Engineering Design: A Review. In: *Journal of Mechanical Design*. eprint: 2110.10863 (cs, stat). (Visited on 12/15/2023).
- Regenwetter, Lyle, Srivastava, Akash, Gutfreund, Dan, and Ahmed, Faez (2023). Beyond Statistical Similarity: Rethinking Metrics for Deep Generative Models in Engineering Design. eprint: 2302.02913.
- Rombach, Robin, Blattmann, Andreas, Lorenz, Dominik, Esser, Patrick, and Ommer, Björn (Apr. 2022). High-Resolution Image Synthesis with Latent Diffusion Models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. arXiv. arXiv: 2112.10752 [cs]. (Visited on 12/15/2023).
- Ruiz, Nataniel, Li, Yuanzhen, Jampani, Varun, Pritch, Yael, Rubinstein, Michael, and Aberman, Kfir (2022). DreamBooth: Fine Tuning Text-to-image Diffusion Models for Subject-Driven Generation. In: *arXiv preprint arxiv:2208.12242*.
- Sabour, Amirmojtaba, Fidler, Sanja, and Kreis, Karsten (2024). Align Your Steps: Optimizing Sampling Schedules in Diffusion Models. arXiv: 2404.14507 [cs.CV].
- Saharia, Chitwan, Chan, William, Chang, Huiwen, Lee, Chris, Ho, Jonathan, Salimans, Tim, Fleet, David, and Norouzi, Mohammad (2022a). Palette: Image-to-Image Diffusion Models. In: *In Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings (SIGGRAPH '22 Conference Proceedings.*

- Saharia, Chitwan, Chan, William, Saxena, Saurabh, Li, Lala, Whang, Jay, Denton, Emily, Ghasemipour, Seyed Kamyar Seyed, Ayan, Burcu Karagol, Mahdavi, S. Sara, Lopes, Rapha Gontijo, Salimans, Tim, Ho, Jonathan, Fleet, David J, and Norouzi, Mohammad (2022b). Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In: *Neural Information Processing Systems (NeurIPS 2022)*.
- Schuhmann, Christoph, Beaumont, Romain, Vencu, Richard, Gordon, Cade, Wightman, Ross, Cherti, Mehdi, Coombes, Theo, Katta, Aarush, Mullis, Clayton, Wortsman, Mitchell, Schramowski, Patrick, Kundurthy, Srivatsa, Crowson, Katherine, Schmidt, Ludwig, Kaczmarczyk, Robert, and Jitsev, Jenia (Oct. 2022). LAION-5B: An Open Large-Scale Dataset for Training next Generation Image-Text Models. arXiv: 2210.08402 [cs]. (Visited on 02/06/2024).
- Singh, Mukul, Cambronero, Jose, Gulwani, Sumit, Le, Vu, and Verbruggen, Gust (2023). Assessing GPT4-V on Structured Reasoning Tasks. eprint: 2312.11524.
- Sohn, Kihyuk, Lee, Honglak, and Yan, Xinchen (2015). Learning Structured Output Representation Using Deep Conditional Generative Models. In: Advances in Neural Information Processing Systems 28 (NIPS 2015).
- Song, Jiaming, Meng, Chenlin, and Ermon, Stefano (Oct. 2022). Denoising Diffusion Implicit Models. In: *International Conference on Learning Representations*. arXiv. arXiv: 2010.02502 [cs]. (Visited on 12/15/2023).
- Sun, Jiao, Fu, Deqing, Hu, Yushi, Wang, Su, Rassin, Royi, Juan, Da-Cheng, Alon, Dana, Herrmann, Charles, Steenkiste, Sjoerd van, Krishna, Ranjay, and Rashtchian, Cyrus (2024). DreamSync: Aligning Text-to-Image Generation with Image Understanding Feedback. In: *Synthetic Data for Computer Vision Workshop @ CVPR 2024*. URL: https://openreview.net/forum?id=cAFgDFprbC.
- Szegedy, Christian, Liu, Wei, Jia, Yangqing, Sermanet, Pierre, Reed, Scott, Anguelov, Dragomir, Erhan, Dumitru, Vanhoucke, Vincent, and Rabinovich, Andrew (Sept. 2014). Going Deeper with Convolutions. arXiv: 1409.4842 [cs]. (Visited on 07/15/2024).
- Tang, Luming, Jia, Menglin, Wang, Qianqian, Phoo, Cheng Perng, and Hariharan, Bharath (Dec. 2023). Emergent Correspondence from Image Diffusion. arXiv: 2306.03881 [cs]. (Visited on 02/28/2024).
- Tumanyan, Narek, Geyer, Michal, Bagon, Shai, and Dekel, Tali (Nov. 2022). Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation. eprint: 2211.12572 (cs). (Visited on 12/15/2023).
- vikhyatk (2024). Moondream1.
- Yang, Binxin, Gu, Shuyang, Zhang, Bo, Zhang, Ting, Chen, Xuejin, Sun, Xiaoyan, Chen, Dong, and Wen, Fang (June 2023). Paint by Example: Exemplar-Based Image Editing With Diffusion Models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 18381–18391.
- Zhang, Lvmin, Rao, Anyi, and Agrawala, Maneesh (Nov. 2023). Adding Conditional Control to Text-to-Image Diffusion Models. In: International Conference on Computer Vision. arXiv. arXiv: 2302.05543 [cs]. (Visited on 12/15/2023).
- Zhou, Luowei, Palangi, Hamid, Zhang, Lei, Hu, Houdong, Corso, Jason J., and Gao, Jianfeng (2019). Unified Vision-Language Pre-Training for Image Captioning and VQA. In: *arXiv preprint arXiv:1909.11059*.

Appendix



A Dataset Distribution

Figure 8. Variance of geometric bicycle layouts before (orange) and after (blue) the filtering of the dataset. The blue bars represent GeoBiked, while the orange bars represent Biked (Regenwetter *et al.* 2021).

B Geometric Feature Detection with Diffusion Hyperfeatures - Experiment Details

When using a single source image, the accuracy of the point detection heavily depends on the source images. An average sample that shows a bicylce of very common style and high similarity with many samples in the dataset produces good mean prediction accuracy (Table IV). Outliers however are not predicted well (see Figure 9, top row). Typical error patterns include uncertainty in the detection of the saddle top and tube intersections. When other bicycle styles are used as source images, the prediction accuracy deteriorates significantly. Using a BMW-style as source, the saddle is not detected for any other style than BMW (Figure 9, middle row). When an uncommon style is used as source, only the wheel centers are accurately detected. All other points are prone to significant inaccuracies (Figure 9, bottom row).

Table IV. Experiment results for a single source image. The metrics are computed as a pixel-wise distance between prediction and ground truth and averaged over all 12 geometry-points. The duration measures the processing time of the subset of 150 samples. Image indices from left to right are 1, 2, 31 and 74.

Source	5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5	50- 100- 100- 100- 100- 100- 100- 100- 1	5 Garde Publis	
MAE	2.429	3.385	12.274	2.710
MSE	33.158	41.635	482.969	41.272



Figure 9. Qualitative Comparison of Error Patterns for a single source image. Top row shows the source image with the highest predicition accuracy.



Figure 10. Qualitative Comparison of Error Patterns using two source images. Top row shows the combination of source images with the best prediction accuracy.



Figure 11. Pixel-wise MAE, MSE and Processing Duration over the number of annotated source images.

C Automatic Generation of Text Descriptions with Vision-Language Models

C.1 GPT-4o System-Prompt

```
1 def construct_prompt(
          length: str,
2
          vibe: str,
3
          style: str,
4
          mode: Literal["im-only", "txt-grounded", "im-txt-grounded"] = "im-only",
5
6) -> str:
      length_map = {
7
           "short": "between 5 and 10 words",
8
          "medium": "between 10 and 20 words",
9
          "long": "between 20 and 40 words",
10
11
      7
      mode_task_map = {
12
          "im-only": "images",
          "txt-grounded": "technical data about bicycles",
14
          "im-txt-grounded": "images and technical data about bicycles contained in
15
      them",
16
      7
17
      image_wrapper = 'Images will be wrapped between <image i></image i> tags.\n'
18
      if 'im' in mode else ''
      bike_wrapper = 'Bike data will be wrapped between <data i></data i> tags.\n'
19
      if 'txt' in mode else ''
20
      description_usage = (
          'You do not have to include all, or any, '
21
          'of the bike data in the description if it does not fit the style or vibe.
22
       ,
          'It is important that the description fits the constraints mentioned above
      . n'
          if "txt" in mode else ''
24
      )
25
26
27
      prompt = (
          "Your task is to create descriptions of bicycles based on {0}. ".format(
28
      mode_task_map[mode])
          + "Each description should fulfill the following constraints: \n"
29
          + "- The length of the provided description should be {0}. \n".format(
30
      length_map[length])
          + "- The descriptions should be {0}. \n".format(vibe)
31
          + "- The descriptions should be in the style of a {0}. \n".format(style)
32
          + "{0}".format(image_wrapper)
33
          + "{0}".format(bike_wrapper)
34
          + "Wrap the resulting bike description in <description i></descriptions i>
35
      tags.\n"
          + "There should be *no* newlines in the descriptions.\n"
36
          + "{0}".format(description_usage)
37
38
          + "The descriptions should be very diverse within the given constraints."
      )
39
40
41
      return prompt
```

C.2 Examples of Generated Description

Table V. Examples of generated text descriptions for different bicycles from the GeoBiked dataset. Descriptions are generated with different configurations of the system prompt and for all three input information combinations.

	Length	Vibe	Style	00	6	OTO		J.J.
Image-only Grounding	Short	Technical	Marketing	Aerodynamic racing bike for speed enthusiasts.	Aerodynamic build with unique wheel design.	Chunky tires for superior terrain stability.	Vintage highwheel, iconic and classic design.	Touring bike with durable pannier rack.
	Long	Technical	Marketing	Featuring an aerodynamic frame and deeprim wheels, this road racing bike is engineered for maximum speed and efficiency.	The ultimate racing bicycle features an aerodynamic carbon frame, integrated dual water bottles, and aero wheels for speed enthusiasts.	Unleash your adrenaline with this BMX bike, featuring extrawide knobby tires for maximum grip and a tough steel frame.	Experience classic charm with this vintage highwheeler bicycle.	Versatile touring bike featuring durable frame, pannier for extra storage, and ergonomic handlebars.
	Short	Technical	Midjourney Prompt	Streamlined aero frame, deeprim wheels, aggressive stance.	Aerodynamic bike with disc wheels.	Offroad tires, durable build, robust frameset.	Highwheeler with large front wheel design.	Touring bike with rear pannier, bottle mount.
	Long	Technical	Midjourney Prompt	Futuristic aero road bike featuring deep dish carbon wheels, an integrated seatpost, and streamlined frame geometry designed for highspeed performance and advanced aerodynamics.	Highperformance racing bicycle with an aero frame.	Rugged allterrain bicycle equipped with oversized tyres for maximum traction, a sturdy black frame, crankset designed for heavyduty pedalling, and a robust fork for enhanced stability on rough surfaces.	Vintage pennyfarthing bicycle with a large front wheel and small rear wheel, black frame, and fixed gear drivetrain.	Touring bike equipped with rear pannier, hydration system, and sturdy frame for extended journeys.
	Short	Casual	Marketing	Speed through with this aerodynamic beast.	Speed demon built for the race track.	Fat tires for mastering tough terrains.	Classic high wheel bike with charm.	Bike with pannier for extra storage.
	Long	Casual	Marketing	"Unleash your speed with this aerodynamic road bike designed for ultimate performance and sleek, futuristic aesthetics.	Embrace topspeed thrill with this aerodynamic bike, built for racing with its triblade wheels and double bottle holders.	Shred through the streets with our rugged BMX bike.	Step back in time with this retro Penny Farthing.	Conquer your daily commute with this versatile bike.
	Short	Casual	Midjourney Prompt	Futuristic road bike, performance meets style.	Aero race bike with sleek wheels.	Shred through the streets with our rugged BMX bike.	Vintage pennyfarthing bike, tall wheel wonder.	Touring bike with pannier for adventures.
	Long	Casual	Midjourney Prompt	Cut through the wind with this sleek, aero bike.	Sleek racing bike equipped with trispoke wheels and dual water bottles, built for speed demons.	Beefy tires and a sturdy frame; this bike's built for offroad adventures and tricks with a bold attitude.	Classic highwheeler with a large front wheel and tiny rear wheel.	Touring bike with a rear pannier, ideal for longdistance adventures.
	Short	Technical	Marketing	Lightweight road warrior with spoked rims.	Trispoke road bike, dual bottle cages.	Highperformance dirt bike for extreme jumps.	Speed and efficiency for road warriors.	Perfect touring bike with rigid fork.
Text-only	Long	Technical	Marketing	Discover unparalleled speed with our Road bike, designed with rigid forks and spoked rims for optimal efficiency on smooth surfaces.	Built for the competitive racer, this road bike boasts trispoke rims and a rigid fork for aerodynamic superiority and speed performance.	Conquer jumps and handle the dirt with precision on our dynamic dirt/jump bike, equipped with robust suspension fork and durable spoked rims.	Experience speed and agility with our Road bike featuring spoked rims and a rigid fork for unmatched performance.	Explore endless trails with our rugged Touring bike featuring spoked rims and a rigid fork, perfect for longdistance adventures.
	Short	Technical	Midjourney Prompt	Spoked rims, rigid fork, road setup.	Street Bike, Trispoke Rims.	Suspension fork, spoked rims, dirt/jump ready.	Spoked wheels road bike with rigid fork.	Touring bike with spoked rims, rigid fork.
	Long	Technical	Midjourney Prompt	Streamlined road bike featuring rigid fork and spoked rims, optimized for speed and agility.	Aerodynamic road bike with trispoke rims and rigid fork, equipped with dual bottle mounts for endurance rides.	Dirt/Jump bike fcaturing spoked rims and a robust suspension fork, designed for durability and performance in jumps.	Precisionengineered road bike with spoked rims and rigid fork, designed for streamlined speed and efficiency on paved surfaces.	Touring bike with spoked wheels, rigid fork, and a bottle mount on the downtube.
Grounding	Short	Casual	Marketing	Speed through the streets with ease!.	Speed and style, perfectly combined!.	Master the jumps and dirt trails!.	Zoom through streets with ease!.	Hit the open road with our Touring bike.
	Long	Casual	Marketing	Hit the road with our super sleek and lightweight road bike, perfect for your next pavement adventure!.	Cruise faster and further on this aerodynamic road bike with stylish trispoke wheels.	Get ready to conquer jumps and trails with this rugged dirt bike.	Hit the roads with speed and precision, perfect for your next big adventure.	Perfect for epic crosscountry adventures.
	Short	Casual	Midjourney Prompt	Speedy road bike for smooth rides.	Aerodynamic road racer.	Perfect for adrenalinepumping jumps!.	Ready to ride the road waves!.	Perfect for your next big adventure.
	Long	Casual	Midjourney Prompt	Ride sleek and fast with this road bike designed for pure speed and agility.	Sophisticated and speedy, this road bike with trispoke rims is ready to roll.	Perfect for dirt trails and jumps, this bike has strong spoked rims and a suspension fork to absorb all the bumps.	Imagine speeding down smooth roads on a sleek, spokedrim, road bike.	Perfect for those long adventures, this touring bike with spoked rims is ready to roll wherever you want to go.
	Short	Technical	Marketing	Streamlined road bike with rigid fork.	Highperformance road bike with trispoke wheels.	Highperformance dirt jump bike.	Classic highwheel, spoked rims, rigid fork.	Touring bike with pannier, perfect for adventures.
	Long	Technical	Marketing	Experience unmatched speed with our sleek Road bike featuring a rigid fork and spoked rims, designed for highperformance road cycling.	Dominate the roads with this aerodynamic bike, equipped with trispoke rims, dual bottle mounts, and a rigid fork for ultimate speed and efficiency.	Robust dirt jumper with spoked rims and suspension fork, designed for highimpact stunts and offroad adventures.	Step back in time with this elegantly designed highwheeler, featuring classic spoked rims and a rigid fork, perfect for enduring vintage rides.	"Perfect for long journeys, this touring bike includes a downtube bottle cage and spoked rims, ensuring reliability and endurance with a rigid fork for stability.
	Short	Technical	Midjourney Prompt	sleek aero road bike design.	Road bike with trispoke rims.	Sturdy jump bike, spoked rims, suspension fork.	Vintage highwheel with rigid fork.	Touring bike with pannier, spoked rims.
Image and Text Grounding	Long	Technical	Midjourney Prompt	Sleek aero road bike, spoked rims, rigid fork, designed for highspeed performance, minimalistic design.	Aerodynamic road bike featuring trispoke wheels, dual bottle mounts, and rigid fork for highspeed performance.	Jump bike, spoked rims, sturdy suspension fork, built for dirt tracks and extreme stunts.	Vintage highwheel bicycle with rigid fork, spoked rims, and elevated seating for a classic riding experience.	Touring bike, rigid fork, downtube bottle cage, equipped with rear pannier, spoked rims.
	Short	Casual	Marketing	Speed and style for road warriors!.	Speed and style in a road package!.	Conquer dirt trails with ease.	Retro charm meets functionality.	Ultimate touring companion for your adventures!.
	Long	Casual	Marketing	Experience speed like never before with this sleek, aerodynamic road bike.	Designed for speed and performance, this road bike is a gamechanger for serious cyclists.	Get ready to catch air and nail tricks with this dirt jump bike, featuring strong spoked rims and a sturdy suspension fork.	Vintage style, ride with elegance!.	Adventure awaits! This touring bike is ready to carry your essentials and take you on epic rides.
	Short	Casual	Midjourney Prompt	Sleek road warrior, raceready speed.	Sleek aero road bike with trispoke wheels.	Bold dirt jumper ready for tricks!.	Vintage charm, high wheel adventure.	Adventureready touring bike with pannier.
	Long	Casual	Midjourney Prompt	Sleek and aerodynamic road bike built for speed; perfect for hitting the pavement and chasing the horizon.	Sleek road bike with trispoke wheels, dual water bottles, and an aerodynamic design for optimal speed and endurance.	This dirt jumper boasts chunky tires, spoked rims, and a suspension fork— built for serious air time and stunts.	Classic pennyfarthing vibes with a mega front wheel and minimalist style, perfect for vintage bike enthusiasts.	Touring bike with a rear pannier, ideal for long adventures.

C.3 Gpt-4o vs. Moondream



Figure 12. Levenshtein distance of unique outputs generated by GPT-40 configurations compared to Moondream. The descriptions generated by Moondream are significantly less diverse.