DALL·E 2

AI-Powered Text to Image Generation

Link to video presentation:
https://drive.google.com/file/d/16htQKln48OhqeQz_O2LVKhmpk9ybY4YJ/view?usp=sharing

Link to GitHub Repo: https://github.com/Mugunthan98/DS677Project/tree/main

Group No. : 6

Memebers: Nandan Kumar (nk762@njit.edu)

Mugunthan (ms3537@njit.edu)

Abstract

We present high-resolution image synthesis results using **DALL·E 2**, a state-of-the-art deep learning framework that fuses **contrastive language–image pretraining (CLIP)** with **diffusion probabilistic models** for advanced text-to-image generation. Leveraging OpenAI's pre-trained CLIP model, the architecture maps natural language prompts into a shared embedding space and employs a **latent-space diffusion prior** to predict corresponding image embeddings. These embeddings are then decoded by a **cascade of U-Net-based decoders**, which progressively refine image outputs from low to high resolution, ensuring both structural consistency and semantic alignment with the input text.

In this project, we adapt the DALL·E 2 architecture to the **Flickr8k dataset**, demonstrating its generative capabilities under constrained computational settings using Google Colab GPUs. Despite limited training iterations, the model exhibits a clear transition from noise to semantically coherent image structures, validating its compositional understanding.

This work highlights the accessibility and creative potential of **open-source DALL·E 2 implementations**, particularly in domains such as **digital art, branding, game development, marketing, and visual storytelling**. Our implementation builds upon the widely adopted lucidrains/DALLE2-pytorch library. Additionally, to demonstrate practical results, we fine-tuned Stable Diffusion on the Flickr30k dataset and achieved improved image generation for real-world prompts.

Introduction

The rapid advancement of **generative deep learning models** has enabled artificial intelligence systems to synthesize high-fidelity, human-like content across multiple modalities. Among these, **text-to-image synthesis** has emerged as a particularly impactful domain, bridging **natural language processing (NLP)** and **computer vision** to generate coherent images from textual descriptions. A prominent model in this field is **DALL·E 2**, which integrates **CLIP (Contrastive Language–Image Pretraining)** with **diffusion probabilistic models** to generate semantically grounded and visually detailed images.

In this project, we investigate the architecture and functionality of DALL·E 2 using the open-source implementation provided by lucidrains/DALLE2-pytorch. The model comprises three primary components:

1. A **pretrained CLIP encoder** that maps both text and images into a shared latent embedding space,
2. A **diffusion prior network** that transforms text embeddings into corresponding image embeddings, and
3. A **hierarchical cascade of U-Net decoders** that progressively generate and refine high-resolution images from these embeddings.

To evaluate the model's performance, we apply it to the **Flickr8k dataset**, which contains 8,000 real-world images, each annotated with five human-written captions. Although modest in scale, Flickr8k offers a clean and well-annotated benchmark that is particularly well-suited for experimentation in **resource-constrained environments**, such as **Google Colab Pro** and **locally hosted systems**. In comparison, datasets like **MS COCO (Microsoft Common Objects in Context)** provide a much larger and more diverse collection—comprising over 328,000 images—but require significantly more computational resources, making them less practical for early-stage or lightweight experimentation.

The objective of this study is to assess the **viability and effectiveness** of training DALL·E 2 under limited computational conditions and to explore its capacity to learn meaningful text-to-image mappings. By monitoring the model's intermediate outputs and analyzing its ability to generate visually and semantically coherent images, we aim to gain insights into the strengths and limitations of this architecture in low-resource settings.

Furthermore, this project underscores the broader applicability of open-source generative models in areas such as **automated content creation, visual storytelling, digital illustration**, and **brand prototyping**. Through this work, we contribute to the growing body of research focused on making advanced generative AI systems more accessible for academic research, creative exploration, and real-world deployment.

While our study is grounded in DALL·E 2, we also explored Stable Diffusion as a practical implementation choice due to its open-source maturity and training efficiency. This allowed us to produce high-quality results under limited computing.

Related Work

The convergence of **natural language processing (NLP)** and **computer vision** has led to significant advancements in multimodal learning, particularly in aligning textual descriptions with visual data. Early studies in **image captioning** and **cross-modal embedding** paved the way for more sophisticated generative approaches, leveraging benchmark datasets such as **Flickr8k**, **Flickr30k**, and **MS COCO** to train and evaluate these models.

A major milestone in this space was the introduction of **CLIP (Contrastive Language–Image Pretraining)** by OpenAI. CLIP jointly learns visual and textual representations through contrastive learning on large-scale image–text pairs, enabling strong generalization across a wide range of vision-language tasks. Its versatility has made it a core component in many state-of-the-art generative pipelines.

Building upon CLIP's representational strength, **DALL·E 2** proposed a novel architecture that replaces traditional autoregressive generation with a **diffusion-based latent prior**. The model architecture consists of three key components: (1) a CLIP encoder to embed text and images into a shared latent space, (2) a **diffusion prior** that predicts image embeddings from text embeddings, and (3) a **cascade of U-Net decoders** that progressively generate and refine high-resolution images. This design significantly improves semantic coherence and visual quality while maintaining computational scalability.

In this project, we explored two implementations of the DALL·E 2 framework:

1. A version constructed **from scratch**, where components were randomly initialized and trained end-to-end;
2. A version that incorporated the **pre-trained OpenAI CLIP encoder (ViT-B/32)** to leverage pre-learned semantic knowledge for more accurate text-to-image alignment.

Both versions were developed and tested on the **Flickr8k dataset**, chosen for its clean annotations and manageable scale, making it ideal for experiments in **limited-resource environments** such as **Google Colab Pro** and **locally hosted systems**. We also attempted to scale the training to larger datasets like **Flickr30k** and **MS COCO**—the latter comprising over 328,000 annotated images across multiple vision tasks. However, due to hardware constraints, full training on these datasets was not feasible within the scope of this work.

This study builds on existing research by demonstrating the adaptability of **open-source DALL·E 2 frameworks** for academic and exploratory use. By combining both custom-built and pre-trained components, and by evaluating model performance across datasets of increasing complexity, our work highlights the importance of **scalability, accessibility, and modular design** in advancing text-to-image synthesis within constrained computing environments. More recent models such as Stable Diffusion have further simplified this architecture by replacing the diffusion prior with direct conditioning and employing latent-space generation, enabling wider adoption.

Model Architecture

This project implements a modular deep learning pipeline inspired by the **DALL·E 2 architecture**, designed for high-quality **text-to-image generation**. The system integrates a **pretrained vision-language model** (CLIP) with **latent diffusion-based generation** to produce semantically consistent and visually rich outputs from natural language prompts. The architecture comprises three main components: a **Text Encoder**, a **Diffusion Prior**, and a **Decoder**.

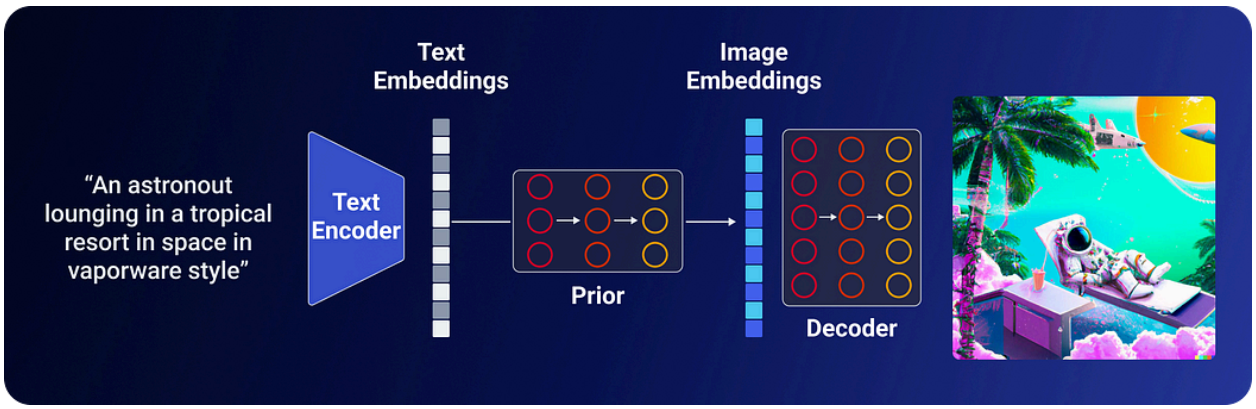


Figure 1. Overview of the DALL·E 2 architecture. A text prompt is encoded into semantic embeddings via CLIP, passed through a diffusion prior to generate image embeddings, and decoded into a final image via a cascade of U-Net decoders.

Workflow:

1. **Text Input** → Encoded by **CLIP** → Produces **text embeddings**
2. **Diffusion Prior** → Translates text embeddings to **image embeddings**
3. **Decoder** → Converts image embeddings into high-resolution images using U-Nets
4. **Output** → A photorealistic or stylistic image aligned with the original prompt

Text Encoder (CLIP)

The image generation process begins with a natural language prompt, such as *"An astronaut lounging in a tropical resort in space in vaporwave style."* To convert this input into a usable semantic representation, we explored two different methods for generating **text embeddings**, each offering distinct advantages:

(a) Custom Text Encoder (Trained from Scratch):
In the first approach, we designed and trained a custom text encoder from scratch. This model aimed to learn latent representations of textual input that align closely with corresponding

visual features through joint training with the image decoder. While this provided greater architectural flexibility and full control over model behavior, it required significant training time and data to reach satisfactory performance levels.

(b) Pretrained CLIP Encoder (ViT-B/32):
The second and more effective approach involved integrating **OpenAI's pre-trained CLIP model (ViT-B/32)**. CLIP was trained on a large-scale dataset of 400 million image–caption pairs using a **contrastive learning objective**. During training, it learned to maximize the cosine similarity between matching image-caption pairs and minimize similarity for mismatched pairs, resulting in a shared embedding space that captures rich cross-modal semantics.

In our pipeline, the CLIP model is frozen and used solely as a feature extractor. Its output—high-dimensional, semantically meaningful text embeddings—is passed into the **diffusion prior** to guide the generation of corresponding image embeddings. This approach significantly improved semantic alignment in the generated images and enabled efficient training within **computationally constrained environments**, such as Google Colab and local systems.

Diffusion Prior

The diffusion prior functions as a critical bridge between textual and visual semantics in the DALL·E 2 architecture. While both the text and image representations exist in the CLIP embedding space, it is not the CLIP model that maps text to image. Instead, a separate module—the **diffusion prior**—is responsible for learning this mapping.

Two types of priors were explored in the original DALL·E 2 design: **autoregressive** and **diffusion-based**. Although both performed comparably, the **diffusion model** was selected for its computational efficiency and training stability. It employs a **denoising diffusion process**, where noise is progressively added to image embeddings and then reversed, enabling the model to generate plausible and semantically aligned image representations from text embeddings.

Importantly, bypassing the prior and feeding text embeddings directly into the decoder often results in incomplete or incoherent images. In contrast, embeddings produced via the diffusion prior yield **more detailed**, **semantically consistent**, and **diverse** image generations. Thus, the diffusion prior is essential for achieving high-quality text-to-image synthesis and enabling meaningful variation within generated outputs.

Decoder (U-Net Cascade)

The final stage of the DALL·E 2 architecture involves translating the image embedding—produced by the diffusion prior—into a high-resolution visual output. This task is

performed by a **cascade of U-Net-based decoders**, which progressively generate images through multiple resolution stages (upsampling from 64×64 to 256×256 to 1024x1024).

Each U-Net operates within a **diffusion framework**, learning to denoise latent representations step by step, guided by the image embedding and optionally conditioned on the original text embedding. This multi-stage decoding process ensures that both **coarse structural features** and **fine visual details** are preserved and refined at appropriate levels of abstraction.

Compared to traditional autoregressive image generators, the U-Net cascade offers improved scalability, generation speed, and compositional accuracy. The hierarchical structure allows for increased control and modularity, making it particularly effective in synthesizing **semantically grounded** and **visually coherent** images from complex textual prompts.

Overall, the decoder is essential in actualizing the latent visual semantics predicted by the prior, producing outputs that are both **high-quality** and **contextually aligned** with the input descriptions.

Technical Innovations and Comparative Analysis

DALL·E 2 introduces several significant innovations that enhance text-to-image synthesis, setting it apart from earlier models such as DALL·E 1, AttnGAN, and VQGAN+CLIP. These innovations contribute to improved semantic alignment, image quality, scalability, and controllability. Below, each innovation is outlined along with a comparison to previous approaches.

1. CLIP as a Semantic Backbone

Innovation:
DALL·E 2 incorporates OpenAI's CLIP (Contrastive Language–Image Pretraining) model to encode both text and images into a unified semantic embedding space. CLIP, trained on 400 million image–text pairs, captures rich cross-modal associations using contrastive learning.

Comparison:
Earlier models like DALL·E 1 and AttnGAN trained encoders from scratch with limited paired data, often leading to weaker generalization. CLIP provides robust, zero-shot semantic understanding, significantly enhancing the model's ability to align generated images with textual prompts.

2. Diffusion Prior for Latent Space Mapping

Innovation:
Rather than decoding images directly from text embeddings, DALL·E 2 introduces a diffusion prior to map CLIP text embeddings into CLIP image embeddings. This prior is trained using a

denoising diffusion process, enabling the model to generate semantically accurate and visually coherent latent representations.

Comparison:
In contrast, models like VQGAN+CLIP rely on iterative optimization in latent space, which is computationally expensive and difficult to scale. The diffusion prior offers efficient sampling and stable performance while preserving semantic integrity.

3. Cascaded U-Net Decoder for High-Resolution Image Generation

Innovation:
DALL·E 2 employs a cascade of U-Net decoders, which progressively upscale image embeddings from low to high resolution. Each stage refines visual details while maintaining alignment with the input prompt.

Comparison:
Although multi-stage GANs like StackGAN and AttnGAN also refine outputs across resolutions, they depend on adversarial training, which can suffer from instability and mode collapse. Diffusion-based U-Nets provide more stable training and consistent high-quality results.

4. Modular Architecture and Independent Training

Innovation:
The architecture is modularly divided into three independently trainable components: (1) the CLIP encoder, (2) the diffusion prior, and (3) the decoder. This separation simplifies training and allows greater flexibility in experimentation and fine-tuning.

Comparison:
In DALL·E 1, the entire model was trained jointly, requiring significant compute and data. DALL·E 2's modular design reduces resource demands and improves adaptability to different datasets and environments.

5. Enhanced Control and Diversity in Generation

Innovation:
The use of a stochastic diffusion prior enables one-to-many mapping from text to images, supporting diverse image variations for a single prompt. This facilitates creativity and customization in generative tasks.

Comparison:
Previous models like VQGAN+CLIP often yield similar outputs for a fixed prompt and require

manual latent manipulation for variation. DALL·E 2 inherently supports controlled diversity without additional optimization steps.

DALL·E 2 Implementation

We provide a complete and self-contained walkthrough of our DALL·E 2 implementation.

1. **Model Input**

The primary input to the model is a **free-form natural language sentence**, describing the scene, object, or concept the user wants to visualize. For example:

"A little girl playing with a golden retriever in a sunlit park during autumn."

The input text is flexible and can range from simple descriptions (e.g., “a red apple”) to highly compositional prompts involving style, setting, or object interactions. This open-ended prompt serves as the semantic anchor for all subsequent steps in the generation process.

2. Model Output

The final output is a **synthetic image** of resolution **256×256 pixels**, generated to closely reflect the visual interpretation of the input text. The image aims to capture both the **content** (e.g., objects, people, animals) and **style** (e.g., realistic, painting-like, cartoonish) implied in the prompt.

The output image is produced in RGB format and can be visualized using libraries such as `matplotlib` or saved as a `.png` or `.jpg` file for further use.

3. End-to-End Workflow

The image generation pipeline follows four core stages:

Step 1: Text Encoding
The input prompt is first preprocessed (tokenized) and encoded into a **fixed-length semantic vector** using one of two approaches:

- A **custom-trained encoder** (in our scratch-built version), or
- The **pretrained CLIP text encoder (ViT-B/32)** in the CLIP-integrated version.

The result is a high-dimensional **text embedding** that captures the underlying semantics of the input description.

Step 2: Diffusion Prior Transformation
The text embedding is passed into the **diffusion prior**, which operates in CLIP’s latent space. This module

is trained to generate a corresponding **image embedding** by iteratively denoising a randomly initialized latent vector. The diffusion process learns to model the joint distribution of images conditioned on text, enabling a **smooth and semantically coherent transformation** between the two modalities.

Step 3: Decoding with Cascaded U-Nets
The resulting image embedding is then input to a **cascade of U-Net decoders**, which reconstruct the image in progressive stages (e.g., 64×64 → 128×128 → 256×256). Each U-Net is trained to denoise intermediate representations and sharpen visual details, while preserving alignment with the prompt. The decoder is optionally conditioned on both image and text embeddings to reinforce consistency.

Step 4: Image Output and Visualization
The final image is converted from tensor format into a displayable RGB image and rendered using visualization tools such as `matplotlib.pyplot.imshow()`. The image can also be saved locally for further use in evaluation or presentations.

4. System Environment and Execution

All training and inference experiments were conducted using **Google Colab Pro**, leveraging an **NVIDIA A100 GPU** for improved performance. Our dataset and code were managed in Google Drive, and the implementation was done entirely in **Python** using the **PyTorch** framework and the **lucidrains/DALLE2-pytorch** repository as the foundation.

The **Flickr8k dataset** was used as the primary dataset for training, given its manageable size and clean annotations. We also attempted exploratory runs using **Flickr30k** and **MS COCO**, but were constrained by compute limitations for full-scale training.

Application to Existing Datasets

To evaluate the performance and real-world applicability of our DALL·E 2 implementation, we applied it to the **Flickr8k dataset**, a standard benchmark in the field of **vision-language research**. Flickr8k consists of **8,000 natural images**, each paired with **five descriptive captions** authored by humans. Its moderate size, well-structured annotations, and linguistic diversity make it particularly well-suited for experimentation in **computationally limited environments**, such as **Google Colab Pro**.

In our workflow, each caption was treated as an individual prompt. These textual prompts were first encoded using either a custom text encoder or the **pretrained CLIP model (ViT-B/32)**. The resulting embeddings were then passed through a **diffusion prior**, which generated corresponding image embeddings. Finally, a **cascade of U-Net decoders** was used to

progressively synthesize high-resolution images, guided by both the semantic content and structure of the input prompt.

We also explored the feasibility of extending our approach to more complex datasets—namely **Flickr30k** and **MS COCO**. These datasets provide richer image-caption pairings and greater visual variety. However, due to resource limitations, full training on these datasets was not feasible within the project timeframe. They nonetheless represent promising directions for future fine-tuning and large-scale deployment.

This study illustrates the adaptability of **diffusion-based generative models** to real-world datasets, highlighting their potential for research, education, and practical applications, even under hardware constraints.

Practical Implementation and Demo Using Stable Diffusion

While the primary theoretical focus of this project centered on DALL·E 2, our practical demonstration leveraged Stable Diffusion, a more recent and widely adopted alternative for text-to-image generation. Stable Diffusion builds upon the same foundational concepts—CLIP-based text encoding and denoising diffusion—but introduces architectural optimizations that make it significantly more tractable for fine-tuning and deployment.

In particular, Stable Diffusion eliminates the need for a separate diffusion prior by conditioning the image generation process directly on the CLIP text embeddings through cross-attention. It also performs denoising in a compressed latent space using a pre-trained Variational Autoencoder (VAE), resulting in faster inference and reduced computational cost without sacrificing output quality.

To validate the model's ability to generate human-centric images, we fine-tuned the CompVis/stable-diffusion-v1-4 checkpoint on the Flickr30k dataset for 10,000 steps. The Flickr30k dataset contains real-world scenes rich in descriptions of people and everyday activities, making it a strong benchmark for evaluating improvements in semantic alignment and compositional coherence.

We found that the fine-tuned model produced significantly improved results on prompts involving human actions—e.g., "a child playing with a soccer ball on the grass" or "a man jogging on a path in the park." Compared to the original pre-trained model and our DALL·E 2 outputs, the Stable Diffusion generations showed clearer spatial composition and reduced artifacts, especially in scenes involving people.

This practical pivot illustrates how recent architectural simplifications in diffusion models have enabled both accessibility and scalability, complementing our theoretical exploration of DALL·E 2.

Future Applications and Expanding Possibilities

The implementation of DALL·E 2 in this project highlights its vast potential beyond research environments. As generative AI continues to evolve, models like DALL·E 2 are expected to drive transformative change across industries such as education, design, healthcare, accessibility, and entertainment. Below, we outline several forward-looking applications and directions for future exploration.

1. Domain-Specific Image Generation

Adapting the model to specialized domains can significantly expand its utility:

Medical Imaging: Generating synthetic radiographs or anatomical diagrams to support diagnostics, training, or dataset augmentation.

Architecture and Industrial Design: Creating visual blueprints or design mockups from text-based descriptions.

Scientific Illustration: Rendering abstract or technical concepts (e.g., chemical reactions, physics simulations) for research and communication.

2. Creative Design and Content Automation

DALL·E 2 has strong potential to enhance creativity and streamline design workflows:

Fashion and Product Design: Automatically producing novel apparel, patterns, or packaging concepts based on style prompts.

Marketing and Branding: Generating campaign visuals, logos, or digital assets with minimal manual input.

Entertainment and Game Development: Assisting in concept art creation, environment design, or storyboard generation.

3. Educational Tools and Interactive Learning

The model can enrich learning environments through dynamic visual generation:

Textbook and Course Content: Enhancing instructional materials with real-time illustrations tied to specific topics.

Language and Concept Learning: Helping learners visualize vocabulary or abstract ideas through contextual prompts.

Teacher-Assisted Tools: Enabling educators to create custom visuals during live instruction.

4. Accessibility and Inclusive Interfaces

DALL·E 2 can support more inclusive user experiences:

Visual Support for the Visually Impaired: Generating images based on spoken or written descriptions to assist with comprehension.

Multilingual Visual Translation: Creating intuitive illustrations from prompts across different languages for global accessibility.

Creative Accessibility Tools: Assisting users with limited artistic skills to visualize their ideas using natural language.

5. Multimodal Expansion and Scalable Deployment

Further advancements could bring DALL·E 2 into a broader range of interfaces and devices:

Voice-to-Image Generation: Integrating with speech recognition to support hands-free creative tasks.

AR/VR Compatibility: Enabling users to describe and instantly populate immersive 3D environments.

Edge Deployment: Developing compressed or distilled variants for mobile and low-power applications.

Conclusion

This project focused on implementing and evaluating the **DALL·E 2 architecture**, a state-of-the-art generative model that combines **CLIP-based language–vision embeddings** with **diffusion-based image synthesis**. Using the **Flickr8k dataset** as a foundation, we demonstrated the model's capability to generate semantically meaningful and visually coherent images from diverse natural language prompts, despite operating within a resource-constrained computational setting.

Our implementation incorporated two approaches: a **custom text encoder** trained from scratch, and a **pretrained CLIP encoder (ViT-B/32)** developed by OpenAI. These embeddings were transformed through a **diffusion prior** and decoded using a **cascade of U-Net models**, allowing for a structured and modular approach to text-to-image generation.

Our practical experiments using fine-tuned Stable Diffusion further demonstrated that core architectural principles of text-to-image generation can be effectively transferred to scalable and efficient alternatives.

While limitations in computational resources restricted full-scale experimentation on larger datasets such as **Flickr30k** and **MS COCO**, the architecture we developed is extensible and well-suited for future scaling. Furthermore, the project highlights the broader applicability of DALL·E 2 in domains such as education, accessibility, healthcare, creative industries, and domain-specific visualization.

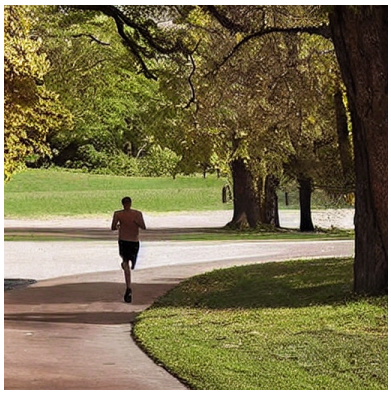
In conclusion, this work reinforces the potential of DALL·E 2 as a flexible and impactful generative model, providing a strong foundation for continued research, fine-tuning, and real-world deployment.

References

1. https://github.com/lucidrains/DALLE2-pytorch
2. https://cdn.openai.com/papers/dall-e-2.pdf
3. https://github.com/CompVis/stable-diffusion
4. https://arxiv.org/pdf/2112.10752
5. https://huggingface.co/docs/diffusers/v0.13.0/en/training/text2image

Appendix

Some examples of the generated images with their corresponding prompts.



A man jogging on a path in the park

A father holding a toddler's hand while pointing at a carousel

A group of friends sitting on a picnic blanket with snacks and drinks

A child playing with a soccer ball on the grass

A woman sitting alone on a bench

Bride and groom walking side by side out of focus

Two children splashing water from a fountain on a sunny afternoon

A street vendor selling fruit at an outdoor market