(1) **更全面的描述**：为了使得预训练的 MiniGPT-4 生成更加全面、更加综合的文本描述，避免不完整、残缺的句子生成。研究者使用构建提示的策略，鼓励基于 Vicuna 的多模态模型生成给定图像的全面描述。具体的提示模板如下所示：

```
###Human: <Img><ImageFeature></Img> Describe this image in detail.
Give as many details as possible. Say everything you see. ###Assistant:
```

其中，*###Human* 和 *###Assistant* 分别代表用户输入和大语言模型的输出。*<Img></Img>* 作为提示符，标记了一张图像输入的起止点。*<ImageFeature>* 代表输入图像在经过视觉编码器和线性投影层后的视觉特征。在这步操作中，一共从 Conceptual Caption 数据集中随机选择了 5,000 张图像，生成对应的、内容更加丰富的文本描述。

(2) **更高质量的描述**：如上一节中所述，由于预训练的 MiniGPT-4 并不能生成高质量的文本描述，仍然存在较多的错误和噪音，例如不连贯的陈述、单词或句子的反复。研究者利用 ChatGPT 强大的语言理解和生成能力，让其作为一个自动化的文本质量评估者，对生成的 5,000 份图像-文本对进行检查。期望通过这步操作，修正文本描述中的语义、语法错误或结构问题。该步操作使用 ChatGPT 自动地改进描述。具体的提示模板如下所示：

```
Fix the error in the given paragraph.
Remove any repeating sentences, meaningless characters, not English sentences, and so on.
Remove unnecessary repetition. Rewrite any incomplete sentences.
Return directly the results without explanation.
Return directly the input paragraph if it is already correct without explanation.
```

在经过 ChatGPT 的评估与改进后，5000 个图像-文本对中最终保留下 3500 对符合要求的高质量数据，用于下一阶段的模型微调。具体的数据格式如图7.11所示，包含基本的图像信息和更加全面的文本描述。

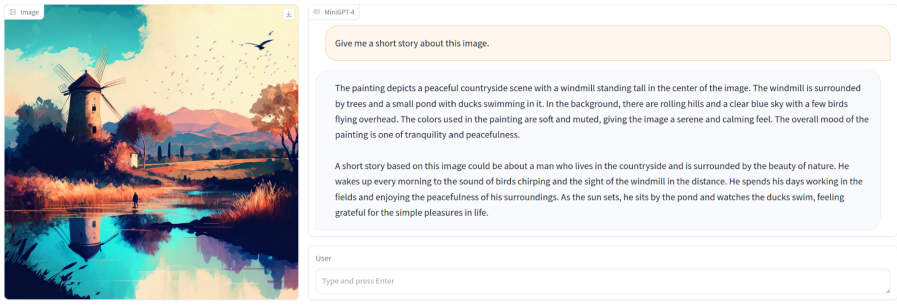

图 7.11　高质量图像-文本数据集示例