## Multimodal Query

CIR Dataset

**Reference Image**

**Moddification Instruction**

*Remove the flowers and made the dog younger and facing the camera with a red collar on.*

## Automated Annotation

Describe the visual details of the reference image in detail.

<Caption>A gray dog with a white beard lies on grass next to a cluster of pink and purple flowers.</Caption>

Explain step-by-step how the modification instruction affects the identified visual elements....

<Reasoning> The flowers **need to be completely removed** from the scene. To make the dog appear younger, its facial features, such as the beard and eyes, **should be altered to look more youthful**. The dog's position needs to be adjusted so that it faces the camera directly. Lastly, a red collar should be added around the dog's neck..... </Reasoning>

Provide a complete description of the resulting target image after applying the instruction.

<Conclusion>A young gray dog wearing a red collar faces the camera directly on a grassy area....</Conclusion>

## Annotation Filtering

InternVL MiMo Keye

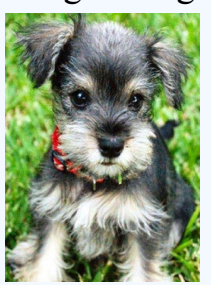
You are an impartial judge whether the "conclusion" text accurately matches the target image. Please assign a score from 1 (Poor match) to 5 (Excellent match) to reflect the quality of alignment. Here is the Scoring Guidelines.....

**Raw Data**

**Avg Score > 2**

**Gold Data**

Target Image



**Comparison**

**Analysis**