A Multi-Modal Image Search Interface Based on CLIP: Implementation and Analysis

1. Dataset Overview

In this project, we constructed a diverse dataset for image retrieval by combining multiple publicly available image sources. The core of the dataset includes:

- **2640 images** from the GroceryStoreDataset, specifically the *Fruits* , *Packages* , and *Vegetables* categories.
- **2973 additional images** , comprising portraits, nature scenes, urban landscapes, and artworks.

This results in a **total of 5613 images** , covering both structured (grocery) and unstructured (natural and artistic) categories. Such diversity allows for evaluating the generalization of multi-modal image retrieval systems across domains.

The dataset was preprocessed using OpenAI's CLIP (ViT-B/32) model to extract feature vectors from both images and texts, enabling cross-modal retrieval.

2. Interface Design Based on the Five-Stage Search Framework

Our image search system was carefully designed to align with the **Five-Stage Search Framework** (Formulation, Initiation, Review, Refinement, Use). Below is how each stage is realized:

2.1 Formulation

- Users can **upload an image** , **input a text description** , or **combine both** as a query.
- A real-time **query preview** is provided within the interface before submission, enabling users to verify their input.

2.2 Initiation

- A dedicated **search button** triggers the retrieval operation.
- The system supports three search modes:
  - Text-based search
  - Image-based search
  - Combined image + text search with adjustable weight

2.3 Review

- After submission, users are shown a **gallery of search results** (retrieved top-k similar images).
- The **total number of returned images** is adjustable via a slider, helping users control result density.

2.4 Refinement

- Users can **adjust retrieval parameters** such as:
  - The number of results shown
  - The image/text feature weight in combined search
- Upon adjusting parameters, the interface automatically updates the result gallery.

2.5 Use

- Users can **download selected images** or **add them to a favorites list** .
- This supports further content curation and reuse.

3. Impact of Different Input Modes on User Experience

Our interface supports two primary query modes:

3.1 Text Input Mode

- Suitable for users with a clear idea of what they are searching for (e.g., “A man in a forest”).
- **Advantages:**
  - Quick to use
  - Flexible in expression
- **Challenges:**
  - Text descriptions may be ambiguous or too general
  - Results depend on CLIP's semantic understanding

3.2 Image Input Mode

- Suitable for users who have a visual reference but cannot describe it accurately.
- **Advantages:**
  - Direct visual matching
  - Useful for non-verbal content (e.g., abstract art)

- **Challenges:**
  - Requires the image to be available beforehand
  - Less flexible if no editing or combination is allowed

3.3 Combined Mode (Image + Text)

- This hybrid input allows **fine-tuning** of intent. For example, uploading an image of a fruit and adding “ripe banana” narrows down the context.
- A **weight slider** lets users control the balance between visual and textual relevance.

4. Making Both Input Modes Equally User-Friendly

To ensure that both input types are equally accessible and intuitive, the following design strategies were applied:

4.1 Unified Search Interface

- All input modes are accessible via a **single dropdown** menu.
- Dynamic form visibility changes based on the chosen mode (e.g., showing image uploader for image mode).

4.2 Preview and Feedback

- The interface provides a **live preview** of the text or uploaded image to confirm user input.
- Querying results are returned quickly, ensuring **immediate feedback** .

4.3 Adaptive Controls

- A **weight adjustment slider** in hybrid mode gives users control over how much emphasis is placed on each modality.
- A result count slider improves accessibility for both novice and advanced users.

4.4 Error Tolerance and Fallback

- If no results are found, fallback messages or alternative suggestions help users adjust queries.
- The system handles file errors and unsupported formats gracefully, enhancing robustness.

5. Conclusion

This project implements a fully functional **multi-modal image retrieval interface** that supports both **textual and visual search** , closely following the Five-Stage Search Framework. The system leverages CLIP to unify vision and language features and integrates an intuitive Gradio-powered UI to optimize user experience.

Through careful handling of both input modes, the system ensures that users—regardless of their preferred querying method—can search efficiently, review results meaningfully, and refine searches to meet specific needs.

This project not only demonstrates the feasibility of cross-modal image search at scale but also highlights the importance of **interface design** in supporting diverse user behaviors.

References

- Radford, A., et al. “Learning Transferable Visual Models From Natural Language Supervision.” *ICML* , 2021.
- Marcus Klasson, GroceryStoreDataset: https://github.com/marcusklasson/GroceryStoreDataset
- Gradio: https://www.gradio.app
- Upstash Vector Index: https://upstash.com