Visual Question Answering (VQA) - Data layer

Datasets — Benchmarks & Sources

VQA v2

- **What it is:** ~265k images (from MS COCO) with 1.1M human-annotated Q&A pairs.
- **Why it matters:** Standard benchmark for VQA; includes open-ended, multiple-choice, and yes/no questions.
- **Quirks:** Class imbalance — yes/no questions dominate (~40%). Annotation style reflects crowdworker biases.
- **Where:** Hugging Face (`vqa`, `visual_question_answering`).

GQA (Graph Question Answering)

- **What it is:** 113k images (from Visual Genome) with ~22M Q&A pairs grounded in scene graphs.
- **Why it matters:** Focuses on compositional reasoning, spatial relations, and multi-hop inference.
- **Quirks:** Longer, more complex questions than VQA v2; sequence length matters.
- **Where:** Hugging Face (`gqa`).

OK-VQA (Outside Knowledge VQA)

- **What it is:** 14k images with 14k Q&A pairs requiring external/common-sense/world knowledge.
- **Why it matters:** Tests models beyond visual recognition — forces multimodal + external knowledge retrieval.
- **Quirks:** Models without knowledge augmentation perform poorly; high variance in question style.
- **Where:** Hugging Face (`ok_vqa`).

VizWiz

- **What it is:** 31k images taken by blind/low-vision people, with Q&A annotations.
- **Why it matters:** Real-world, noisy data; accessibility-driven benchmark.
- **Quirks:** Images often blurry, poorly lit; many unanswerable questions.
- **Where:** Hugging Face (`vizwiz`).

TextVQA

- **What it is:** 28k images from OpenImages with 45k Q&A pairs that require OCR.
- **Why it matters:** Benchmarks OCR + visual reasoning for VQA (e.g., reading street signs, labels).
- **Quirks:** Strong OCR dependency; fails if text extraction is weak.
- **Where:** Hugging Face (`textvqa`).

Preprocessing (what to do and why)

Resizing

We standardize image size so they can batch efficiently through CNNs/ViTs.

- **Train:** `RandomResizedCrop(224)` → introduces scale/translation invariance.
- **Eval:** `Resize(256) + CenterCrop(224)` → consistent input resolution.

Normalization

We adjust pixel values so they're centered and scaled, making training stable.

- **ImageNet stats:** `Normalize(mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225])` → Matches expectations of pretrained vision backbones (ResNet, ViT).
- **From scratch:** `Standardize per dataset` → Only needed if no pretrained weights.

Tokenization (for Questions)

We convert natural-language questions into token IDs for transformers.

- **HF Tokenizers:** `AutoTokenizer.from_pretrained("bert-base-uncased")` → Common baseline.
- **Seq length:** Pad/truncate to 32–64 tokens (most VQA questions are short).

Feature Extraction (optional, for hybrid pipelines)

We can pre-extract visual features to save compute.

- **Faster R-CNN / DETR:** Object-level features (e.g., 36 region proposals per image).
- **CLIP-ViT:** Global + patch embeddings.

Dataloading tips

We prepare the dataset so training is fast, reproducible, and efficient.

- **Prefetch & pin memory:** `DataLoader(pin_memory=True, prefetch_factor>1)` → Keeps GPU fully utilized.
- **Worker init functions:** `worker_init_fn=seed_all` → Ensures reproducible augmentations.

- **Deterministic validation:** Fixed transforms (`Resize + CenterCrop + Normalize`) → Stable evaluation across runs.
