**Introduction**

**Problem**
- Given a reference image and user text as input query, we consider to retrieve new images that resemble the reference image while changing certain aspects as specified by text.
- The text can be given as attribute or natural language.

**Main Challenges**
- simultaneously preserve and transform the visual content in accordance with the text feedback
- learn a composite representation that jointly encapsulate visual and textual contents from coarse to fine-grain

**Main Idea**
- Visiolinguistic Attention Learning (VAL)
  - model architecture
    - composite transformers at multi-level
    - attentional transformation and preservation
    - fuse vision and language features via attention learning at varying representation depths.
  - learning objective
    - hierarchical matching
      - align with the target visual and textual representations in a two-level hierarchical space

**Proposed Approach**

**Methodology**

**Experiments**

**Experiments on three benchmark datasets**

**Qualitative results**

**Table 1. Fashion200k**

**Table 2. Shoes**

**Table 3. FashionIQ**

**Reference**

[1] Xiaokai Guo, Hui Wu, Yupeng Gao, Steven Rennies, and Rogerio Feris. The fashion iq dataset: Retrieving images by combining side information and natural language feedback. ICCVW’19

[2] Nam Yh, Lu Jiang, Chen Sun, Kevin Murphy, Li-Ba Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval - an empirical odyssey. CVPR19