This is a preprint of Chapter 5 in the upcoming book *Recommendation with Generative Models* by Yashar Deldjoo, Zhankui He, Julian McAuley, Anton Korikov, Scott Sanner, Arnau Ramisa, Rene Vidal, Mahesh Sathiamoorthy, Atoosa Kasrizadeh, Silvia Milano, and Francesco Ricci.

5

Multi-modal Generative Models in Recommendation System

Chapter Authors[1]**:** Arnau Ramisa, Rene Vidal

ABSTRACT

The recommendation systems discussed so far typically limit user inputs to text strings or behavior signals such as clicks and purchases, and system outputs to a list of products sorted by relevance. With the advent of generative AI, users have come to expect richer levels of interactions. In visual search, for example, a user may provide a picture of their desired product along with a natural language modification of the content of the picture (e.g., a dress like the one shown in the picture but in red color). Moreover, users may want to better understand the recommendations they receive by visualizing how the product fits their use case, e.g., with a representation of how a garment might look on them, or how a furniture item might look in their room. Such advanced levels of interaction require recommendation systems that are able to discover both shared and complementary information about the product across modalities, and visualize

[1]This work does not relate to the authors positions at Amazon.

the product in a realistic and informative way. However, existing systems often treat multiple modalities independently: text search is usually done by comparing the user query to product titles and descriptions, while visual search is typically done by comparing an image provided by the customer to product images. We argue that future recommendation systems will benefit from a multi-modal understanding of the products that leverages the rich information retailers have about both customers and products to come up with the best recommendations.

In this chapter we discuss recommendation systems that use multiple data modalities simultaneously. As we shall see, a key challenge in developing multimodal generative models is to ensure that the features extracted from each modality are adequately *aligned* across modalities, i.e., mapped to nearby points in the embedding space. Since the problem of jointly learning a generative model for each modality and their alignment is extremely difficult (Chen *et al.*, 2020a), a common approach is to use contrastive learning methods to approximately align the modalities before learning a multimodal generative model. Therefore, in this chapter we will review both contrastive and generative approaches to multimodal recommendation. More specifically, in Section 5.1 we will provide a brief introduction to multimodal recommendation systems, in Section 5.2 we will review contrastive approaches to multimodal recommendation, and in Section 5.3 we will discuss generative approaches. Finally, in Section 5.4 we will overview various applications of multimodal recommendation systems. Throughout the chapter, we will center the discussion around vision and language models due to the larger volume of work for these two modalities, but we note there is a growing literature of generative recommendation systems that combine other modalities such as audio and text (Vyas *et al.*, 2023), video and audio (Ruan *et al.*, 2023),

or even more than two modalities (Wu *et al.*, 2023b).

5.1 Introduction to Multimodal Recommendation Systems

5.1.1 Why do we need multimodal recommendation systems?

Retailers have a lot of information about their customers and the items they sell, including purchase history, customer interactions, product descriptions, product images and videos, and customer reviews. However, existing recommendation systems typically process each data source independently and then combine the recommendation results. For example, text search is typically done by comparing a short user query to product title, descriptions and reviews, while visual search is typically done by comparing an image provided by the customer to product images. Both search approaches produce a list of products sorted by relevance, and current "multimodal" systems simply fuse unimodal relevance scores to produce a single list of products from both modalities. In practice, there are many use cases in which such a "late fusion" approach may be insufficient for satisfying the needs of the user.

One such use case, known as the *cold start problem*, occurs when new users start using the system, or new products are added to the catalog, hence user behavioral data cannot be leveraged to recommend new products to existing users or existing products to new users. To alleviate this problem, it is useful to gather diverse information about the items so that preference information can be transferred from existing products or users to new ones. To this end, models that combine information from multiple modalities offer a unique advantage. For example, if a store receives a new product (e.g., a dress), but no purchases have been made yet, we can use the visual similarities between the new dress and existing ones in the store to determine which customers could be interested in it.

Another use case occurs when different modalities are necessary to understand the user request. For example, to answer the request "best metal and glass black coffee table under $300 for my living room", the system would need not only the text query but also an image of the

customer's living room in order to find a table that best matches the room. Moreover, answering this customer's question requires reasoning about the appearance and shape of the item in context with the shape and appearance of many other objects, as well as limiting the search by price, which cannot be achieved by searching with either the text or image independently. Other examples of multimodal requests include an image or audio of the desired item together with modification instructions in text (e.g., a dress like the one in the picture but in red, a song like the sound clip provided but in acoustic), or a complementary related product (e.g., a kickstand for the bicycle in the picture, or other movies from the actress talking in the video clip).

A third use case where multimodal understanding becomes crucial is when considering more complex recommendation systems, like those featuring virtual try-on capabilities, or intelligent conversational shopping assistants (Mehta, 2024; Templeton, 2024). To be effective, AI shopping assistants will need to be able to understand the context of previous interactions in the conversation history. Let's consider the example of a customer looking for a complete outfit he is planning to wear during the summer in Cairo, to attend the wedding of a friend with traditional tastes. An AI shopping assistant interacting with the customer will have to resort to visual cues to recommend products compatible as an outfit, as well as other customer preferences expressed earlier, the climate in Cairo during the summer, and cultural or dress code norms.

5.1.2 Key challenges in designing multimodal recommendation systems

The development of multimodal recommendation systems faces several challenges.

- First, combining different data modalities to improve recommendation results is not simple. Existing systems learn joint representations that capture information that is shared across modalities (e.g., the text query refers to a visual attribute of the product that is visible in the image), but they ignore complementary aspects that could benefit recommendations (Guo *et al.*, 2019); e.g., the text mentions inside pockets not visible in the picture, or the im-

age contains texture patterns that are hard to describe precisely in text. Therefore, when learning multimodal representations it is important to ensure adequate alignment of the aspects that need to be aligned, while leaving some flexibility to capture complementary information across modalities as well. In general we will want the modalities to compensate for one another and result in a more complete joint representation.

- Second, collecting aligned data from multiple modalities to train multimodal recommender systems is significantly more difficult than collecting data for individual data modalities. For example, in the unimodal case one can define positive pairs for contrastive learning via data augmentation, but in the multimodal case such positive pairs often need to be annotated (see Section 5.2). In practice, existing annotations may be incomplete for some modalities (Rahate *et al.*, 2022). For example, visual search with text modification would require examples of an input image, the textual modification, and the modified image, but typically only two of the three are available, e.g., image-caption pairs.

- Third, learning a latent space that can be used for generative tasks is often harder than for discriminative tasks, as it typically requires larger datasets and computational resources to be able to adequately learn the data distribution by using more complex losses (Chen *et al.*, 2020a). This challenge is further exacerbated in the case of multimodal data because we need to not only learn a latent representation for each modality but also ensure that these latent representations are adequately aligned.

Despite these challenges, multimodal generative models are a promising technology for improving recommendation systems. Indeed, recent literature shows tremendous advances on the necessary components to achieve effective multimodal generative models for recommender systems, including 1) the use of LLMs and diffusion models to generate synthetic data for labeling purposes (Brooks *et al.*, 2023; Rosenbaum *et al.*, 2022; Nguyen *et al.*, 2024), 2) high quality unimodal encoders and decoders (He *et al.*, 2022; Kirillov *et al.*, 2023), 3) better techniques

for aligning the latent spaces from multiple modalities into a shared one (Radford *et al.*, 2021; Li *et al.*, 2022; Girdhar *et al.*, 2023), 4) efficient re-parametrizations and training algorithms (Jang *et al.*, 2016), and 5) techniques to inject structure to the learned latent space to make the problem tractable (Croitoru *et al.*, 2023; Yang *et al.*, 2023b). Once trained, generative recommender systems are more versatile, and can produce better recommendations in more general, open ended tasks.

5.1.3 Multimodal recommendation systems covered in this chapter

In the remainder of this chapter, we will review both contrastive and generative approaches to multimodal recommendation. In Section 5.2 we will review contrastive approaches, such as CLIP, which learns to map each modality to a common latent space in which the modalities are approximately aligned. In Section 5.3 we will discuss generative approaches, such as ContrastVAE, which learns a probabilistic embedding from each modality to a common latent space where modalities are approximately aligned, and DALL-E 2, Stable Diffusion, LLAVA and multimodal LLMs, which learn to generate image recommendations given an input text prompt.

5.2 Contrastive Multimodal Recommendation Systems

As discussed in Chapter 4.3.1, many recommendation approaches like Du *et al.*, 2022 rely on learning an embedding of the data such that similar items are close to each other in the embedded space. In the case of multimodal data, a natural approach to learning a *multimodal embedding* would be to learn one embedding per modality, as done by (He *et al.*, 2020; Grill *et al.*, 2020; Chen *et al.*, 2020b; Caron *et al.*, 2021) for images, or (Saeed *et al.*, 2021; Won *et al.*, 2020; Wang *et al.*, 2022a) for audio, and then concatenate such *unimodal embeddings*. Such an approach is adequate when different modalities capture complementary aspects of an item. However, as discussed in Section 5.1.2, when different modalities capture related aspects of an item, unimodal embeddings need to be adequately aligned to ensure that similar items are close to each other in the multimodal embedded space. For example, to ensure

that the embedding of a textual description of a product is close to the embedding of an image of the same product we need to learn both embeddings with that constraint in mind, which often requires large amounts of aligned training data (e.g., text-image pairs). One way to address this challenge is to first learn an alignment between data modalities and then learn a generative model on *aligned* representations. Hence, in this section we will focus on the problem of learning aligned representations across multiple modalities.

A popular approach to learning aligned representations is contrastive learning (Gutmann and Hyvärinen, 2010) which, for a pair of data points from different modalities, minimizes a loss that encourages their embeddings to be close when the points are similar (positive pairs), and far when the points are very different (negative pairs). In the single modality setting, positive pairs are generated by simply altering one sample (e.g., slightly shifting the image, flipping the image, transforming it to grayscale). In the multimodal setting, however, it is hard to generate a corresponding positive pair in the other modality via simple augmentation strategies. Instead, positive pairs are typically obtained by labeling similar pairs in a coarse-grained or fine-grained manner. Coarse-grained labels (e.g., a pair of an image and a caption) are easier to obtain, but they may not be sufficiently discriminative. Fine-grained labels (e.g., a bounding box for each object in the image and the corresponding word in the caption) are harder to obtain, but they provide more detailed correspondences between image regions and words in the caption.

5.2.1 Contrastive Language-Image Pre-training (CLIP)

Contrastive Language-Image Pre-training (CLIP) (Radford *et al.*, 2021) is one of the most popular contrastive learning approaches to multimodal pre-training. The main idea behind CLIP is that coarse labels in natural language have a sufficient degree of supervision to enable the learning of general concepts, while being much easier to scale using internet data. Indeed, the authors of CLIP found that trying to predict the exact words, as previous works had done, led to very hard training objectives that converged very slowly, due to the variety of ways in which the same information can be conveyed. Therefore, they proposed

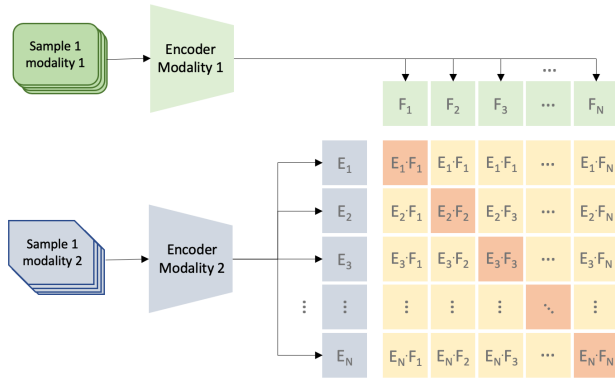

Figure 5.1: Contrastive pre-training used to train models such as CLIP. For each minibatch, the positive (diagonal) and negative (off-diagonal) pairs are used to compute the loss.

to use coarse labels, i.e., to pair an entire image with a caption.

Figure 5.1 shows CLIP's model architecture, which consists of two towers, an image encoder and a text encoder, that project an input image-text pair to a shared embedding space. Semantically equivalent image-text pairs should be projected to the same point in the embedding space, and unrelated image-text pairs should be projected to far apart points. This is achieved by computing the cosine similarity for all possible image-text pairs in a training minibatch, and applying a symmetric cross-entropy loss over the rows and columns of the similarity matrix.

To be effective, CLIP was trained on a large dataset of 400 million image and text pairs, obtained from downloading images and their associated *alt-text* (text to be displayed in place of an image that fails to load) from the Internet. The dataset was curated to guarantee coverage of concepts by balancing the word occurrences, and filtering with methods like making sure each text included at least one word from a pre-defined list obtained from Wikipedia data to remove noisy or irrelevant image-text pairs. While aligning crawled Internet images and their alt-text is bound to find many irrelevant or misleading examples, dataset curation techniques to improve the quality of training samples (Cao *et al.*, 2023; Fan *et al.*, 2023), or to reduce harmful or undesired examples (Bansal *et al.*, 2023; Yu *et al.*, 2023) have proven useful to improve results.

Furthermore, scaling up the datasets to billion-scale (Schuhmann *et al.*, 2022) has shown that noisy examples not previously removed can be cancelled by overwhelming numbers of positive ones, resulting in better overall performance Jia *et al.* (2021).

The simple idea behind CLIP demonstrated to scale very well and achieved state-of-the-art in many zero-shot benchmarks. For example, it obtained impressive zero-shot classification and retrieval results (Novack *et al.*, 2023; Baldrati *et al.*, 2023; Hendriksen *et al.*, 2022), and has been successfully fine-tuned to a multitude of tasks, such as object detection (Gu *et al.*, 2021), semantic segmentation (Zhou *et al.*, 2023b) or action recognition (Huang *et al.*, 2024). The same contrastive alignment objective has also been used between other modalities, including audio and images (Cheng *et al.*, 2020), tables and images (Hager *et al.*, 2023), tables and medical images (Huang, 2023), and with multiple modalities at the same time (Girdhar *et al.*, 2023). The datasets used for pre-training these models are typically composed of data scrapped from the Internet (e.g., pairs of images and alt-text), generated as a byproduct of another process (e.g., e-commerce purchases (Chen *et al.*, 2023b), robot sensor logs (Huang *et al.*, 2021)), or automatically generated by an existing ML model (e.g. speech in audio or video (Zhang *et al.*, 2021), human poses in images computed with OpenPose (Cao *et al.*, 2019)).

The generalization ability of CLIP and similar Vision-Language Models (VLM) greatly benefited from scaling the training in model size, batch size, and dataset size (Pham *et al.*, 2023; Cherti *et al.*, 2023). Researchers have also studied how preferring adaptation of the text branch over the language branch affected results (Zhai *et al.*, 2022). Furthermore, many approaches have been proposed to improve the semantic accuracy of the resulting models (Li *et al.*, 2023a), such as loss functions to improve the image and text encoders (He *et al.*, 2022; Shen *et al.*, 2022), or to encourage desirable properties such as multilanguage understanding, interpretability and fairness in the embedding space (Chen *et al.*, 2023a; Carlsson *et al.*, 2022; Dehdashtian *et al.*, 2024). Other interesting improvements include training better encoders with additional loses like image masking (He *et al.*, 2022) and Triple Contrastive Learning (Yang *et al.*, 2022), or enhancing the text with Wikipedia definitions of entities (Shen *et al.*, 2022).

5.2.2 Other Contrastive Pre-training Approaches

Other approaches have looked into novel architecture designs and novel losses to further improve results. Align BEfore Fuse (ALBEF) (Li *et al.*, 2021), for example, uses a multimodal encoder to combine the text and image embeddings generated from the unimodal encoders, and propose two additional objectives to pre-train a model in addition to the Image-text contrastive (ITC) learning: masked language modeling (MLM) to predict masked words on the unimodal text encoder, and image-text matching (ITM) to classify if a pair of image and text match or not. The authors also introduce *momentum distillation*, where a moving average version of the model weights provides pseudo-labels in order to compensate for the potentially incomplete, or wrong, text descriptions in the noisy web training data. Using their proposed architecture and training objectives, ALBEF obtains better results than CLIP in several zero-shot and fine-tuned multimodal benchmarks, despite using orders of magnitude less images for pre-training. In a subsequent work, Li *et al.* (2022) replace the multimodal encoder by cross-attention layers to the text tower to model vision-language interactions, and replace the MLM loss by a Language Modelling (LM) loss that trains the model to maximize the likelihood of a generated caption given an image.

Finally, other works explore how to bring more modalities into alignment. Girdhar *et al.* (2023) propose ImageBind, an approach to learn an aligned embedding across six different modalities, including text, audio, image, depth, thermal and Inertial Measurement Unit (IMU) data. Instead of requiring paired data for all modalities, they only rely on readily available paired data between image and other modalities (e.g., web scale text-image data, audio for a video clip or depth in RGBD images). All modality encoders use transformer networks and the joint model is learned using the InfoNCE loss.

These contrastive multimodal models can then be used in multimodal recommendation systems such as (Sevegnani *et al.*, 2022; Alpay *et al.*, 2023; Wu *et al.*, 2023b). They are also used to initialize the weights of generative multimodal systems, that will make the generative training much more tractable.

5.3 Generative Multimodal Recommendation Systems

Despite their advantages, purely contrastive recommendation systems often suffer from data sparsity and data uncertainty (Wang *et al.*, 2022c; Lin *et al.*, 2023b). For example, users may provide reviews for very few items and some of them may have errors. Generative models address these issues by imposing structure on the data generation process, e.g., by using latent variable models, and by adequately modeling uncertainty. Moreover, generative models allow for more complex recommendations, e.g., those involving image generation.

In this section, we will survey generative recommendation systems that utilize multiple modalities in order to better understand the user, or provide the recommendations. Depending on how the generative models are designed and learned, we will distinguish between three types of models: Generative Adversarial Networks (see Section 5.3.1), Variational AutoEncoders (see Section 5.3.2), and Diffusion Models (see Section 5.3.3). All these three types of models posit the existence of a latent variable Z (continuous or discrete) such that the distribution of the data X (e.g., image and text) can be written as

$$p(X) = \int p(X \mid Z)p(Z)dZ. \tag{5.1}$$

The main differences among these models are how the prior $p(Z)$ and posterior $p(X \mid Z)$ are defined and parametrized with deep networks, and what losses are used to learn the network weights from data. The following subsections describe each one of these models in more detail, how network architectures are modified to accomodate multimodal data, and how these models are used for building recommendation systems.

5.3.1 Generative Adversarial Networks for Multimodal Recommendation

Proposed by Goodfellow *et al.* (2014), Generative Adversarial Networks (GANs) are an innovative approach to learning a distribution from multimodal data. GANs have been used in various recommender systems, including collaborative filtering (Wei *et al.*, 2023) and content-based retrieval (Tautkute and Trzcinski, 2021). In this subsection, we will briefly

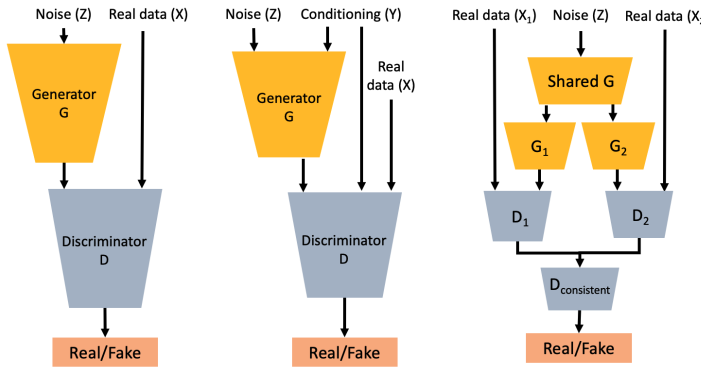

Figure 5.2: Generative Adversarial Networks (GANs) are composed of a generator G that generates a data point (e.g. an image) from a latent variable Z, and a discriminator D which tries to determine if a data point is fake (synthesized by the generator) or real. Left: Standard GAN architecture for unconditional generation. Center: Conditional GAN architecture. Right: Multimodal GAN architecture.

summarize the basic formulation of GANs for unimodal data, show how it can be extended to multimodal data, and discuss adaptations of GANs for collaborative filtering and content-based retrieval.

Unimodal GANs As discussed in Section 3.5.5, GANs are a class of latent variable models in which a generator maps a latent variable Z to a sample data point X, while a discriminator D decides whether its input is real or generated (see Fig. 5.2 left). More specifically, GANs learn a probability distribution as in Eq. (5.1), where $p(Z)$ denotes the prior which is typically assumed to be a standard Gaussian (continuous) or categorical (discrete) and $p(X \mid Z)$ denotes the posterior. What is unique in GANs is that the posterior $p(X \mid Z)$ is not modeled explicitly (e.g., as a Gaussian). Instead, the posterior is represented by a generator $G : \mathcal{Z} \to \mathcal{X}$ that produces samples $G(Z)$ from $p(X \mid Z)$ without having to represent $p(X \mid Z)$. Then, a second component of a GAN is the discriminator $D : \mathcal{X} \to \{0, 1\}$, which is designed to discriminate real from generated images, i.e., $D(X) = 1$ if X is real and $D(X) = 0$ if X is generated. The generator and discriminator are then jointly learned

from samples of $p(X)$ by optimizing a min-max objective

$$\min_G \max_D \mathbb{E}_{X\sim p(X)}[\log D(X)] + \mathbb{E}_{Z\sim p(Z)}[\log(1 - D(G(Z))] \tag{5.2}$$

in which the generator G tries to generate samples that fool the discriminator D, while D tries to discriminate between real samples $X \sim P(X)$ and generated samples $G(Z)$, $Z \sim P(Z)$.

One advantage of GANs is that sampling is straightforward: all we need to do is to sample Z (e.g., categorical or standard Gaussian) and pass it through the generator to produce X. Another advantage is that, in the ideal case in which G and D have infinite capacity, one can show that the optimal discriminator D^* can't tell true from generated (i.e., $D^*(X) = 1/2$) and the optimal generator G^* is such that the distribution of the generated data $G^*(Z)$ matches the distribution of the true data X. In practice, D and G are parametrized with neural networks, and the expectation in the min-max objective is computed as the average over samples. As a consequence, while there is no guarantee that GANs learn the true distribution of the data, in the case of images it has been empirically shown that GANs produce high quality generations.

Despite these advantages, GANs also suffer from some limitations. One of them is the issue of mode collapse, which happens when the generator produces samples that are not representative of the full data distribution, such as generating only the most likely outputs, or a specific output that fools the discriminator (Zhang *et al.*, 2019). GANs also suffer from training instabilities due to the nature of the min-max objective optimized by the generator and discriminator networks. For example, (Arjovsky *et al.*, 2017) show that a small change in one network leads to major adjustments in the other, which can result in destabilizing the learning process and failing to converge. Moreover, gradient vanishing problems happen when one network dominates the other, e.g., when the discriminator becomes very accurate and produces a loss with little gradient information for the generator (Su, 2018; Chakraborty *et al.*, 2024).

Multimodal GANs The vanilla GAN formulation discussed so far assumes that X is generic, i.e., X can be unimodal or multimodal. In principle, we could use such a vanilla formulation to learn generative

models for multimodal data. However, doing so may require collecting, annotating and aligning very large datasets and using them to train a very complex multimodal generator. In practice, it may be preferable to design specialized models that leverage existing unimodal generators, such as models that can generate one modality conditioned on another, or models that can ensure adequate alignment across modalities.

Conditional GANs (see Fig. 5.2 center) generate data for data modality X conditioned on another modality Y, such as the product type, a textual description of a product, an image mask, etc. In this case, the goal is to learn a conditional model of the form

$$p(X \mid Y) = \int p(X \mid Z, Y)p(Z)dZ. \tag{5.3}$$

To model $p(X \mid Z, Y)$, the generator must take both Z and Y as inputs to generate samples $G(Z, Y)$. Likewise, the discriminator $D(X, Y)$ must also depend on the conditioning variable Y. Different modalities can be used for the condition; examples are class-conditioning (Mirza and Osindero, 2014), conditioning on an input image (Isola *et al.*, 2017), or using a latent code vector (Chen *et al.*, 2016). Huang *et al.* (2022) proposed an approach to allow conditioning on multiple input modalities (e.g., text, sketch, segmentation mask) to generate new images. This allowed very fine-grained control of the generated image layout and content. Ziegler *et al.* (2022) also use conditioning on multi-modal clinical tabular data for the generation of realistic 3D medical images.

Alternatively, we may want to generate multiple data modalities. For the sake of simplicity, assume that the data is composed of two modalities $X = (X^1, X^2)$. We can design a multimodal generator that leverages unimodal generators by assuming that X^1 and X^2 are conditionally independent given Z. Under these assumptions, the model in Eq. (5.1) factorizes as the product of two unimodal models because

$$p(X^1, X^2) = \int p(X^1, X^2 \mid Z)p(Z)dZ = \int p(X^1 \mid Z)p(X^2 \mid Z)p(Z)dZ. \tag{5.4}$$

Therefore, we can use one generator per modality, $X^1 = G^1(Z)$ and $X^2 = G^2(Z)$, to represent $p(X^1 \mid Z)$ and $p(X^2 \mid Z)$, respectively. Note, however, that the latent representation Z must be shared to ensure

alignment across modalities. Alternatively, we may want the generators G_1 and G_2 to have a shared backbone that then splits into separate branches for each modality (see Fig. 5.2 top right). For example, Zhu *et al.* (2024) use a StyleGAN backbone with three modality specific branches.

Regarding the design of multimodal discriminators, we note that the discriminator should take generated data for both modalities and compare it with the true data for both modalities. To leverage pre-trained discriminators for each specific modality, say D_1 and D_2, we could simply fuse the predictions of unimodal discriminators. Alternatively, we could fuse intermediate features from unimodal discriminators and have a simple discriminator $D_{consistent}$ predict whether the data is real or fake from the fused representation of both modalities (see Fig. 5.2 bottom right). Zhu *et al.* (2024) also use two types of discriminators: *fidelity discriminators* are unimodal discriminators that assess the quality of an individual data modality, while *consistency discriminators* judge whether two modalities are consistent with each other.

Multimodal GANs for collaborative filtering As discussed before, a natural approach to building multimodal recommendation systems is to incorporate multiple modalities when learning a latent representation of items and/or users. However, existing multimodal representation learning methods lack robustness to scarce labels for user-item interactions. Self-supervised learning methods address this problem by exploiting supervisory signals in unlabeled data, e.g., by using data augmentation. However, a key challenge is generating augmentations that are consistent across multiple modalities. Recent work (Wei *et al.*, 2023) proposes an adversarial multi-modal self-supervised learning paradigm in which a generator proposes collaborative relations which are then vetted by a discriminator. In addition, Wei *et al.* (2023) propose a cross-modal contrastive learning framework for preserving inter-modal semantic commonality and user preference diversity. On the other hand, GANs have been used to model and improve user-item interaction data. For example, Gao *et al.* (2021) review several works that use GANs to mitigate noise and perform informative sample selection in user preferences data, and to synthesize new samples through data augmentation.

Multimodal GANs for fashion recommendation Due to its visual nature, GANs have found an application area in fashion-related tasks such as compatible outfit generation, virtual try-on, or product search.

Liu *et al.* (2021) tackle clothing compatibility learning. Given an image of a clothing product, and a target product category name, the proposed method generates an image of a compatible item with a GAN. A compatibility matrix representing a style space is used to condition the GAN and make sure the generated item is compatible with the input one. The style space is learnt using triplets of anchor with compatible and incompatible clothing items, and additional losses for feature matching, reconstruction, and a discriminator loss. Zhou *et al.* (2023a) propose a method to generate multiple options for compatible clothing simultaneously, with attention to diversity. They use a style embedding discriminator to provide supervision to the generator through a binary real/fake classification loss, and a compatibility discriminator that uses a contrastive loss. They also include a diversity loss to ensure variety in the generated items.

For virtual try-on, an input garment image, and a target image with a person onto which the garment has to be placed are used. Generators often use several modalities derived from the target image, such as the person mask and a pose image, to condition the generation. For example, Liu *et al.* (2019) take conditional and reference images and transfer the clothing from the person in the conditional image to the one in the reference image. For that, they use a pose map, segmentation map, mask map and head map, derived from the input images. They combine three generators and a discriminator to have a single system for clothing transfer. Similarly, Pandey and Savakis (2020) use segmentation masks, pose estimation and clothing parsing (i.e., detecting all clothing in a picture) to transfer a reference garment in a white background image to a person in a model image. Their proposed system combines multiple tasks previously done by different networks into a single architecture.

Tautkute and Trzcinski (2021) use GANs for query expansion (i.e., augmenting or reformulating the user query to improve retrieval results) on a multimodal fashion product retrieval scenario. Instead of combining or fusing the text and image representations, as is commonly done in multimodal search, they generate an image of the desired

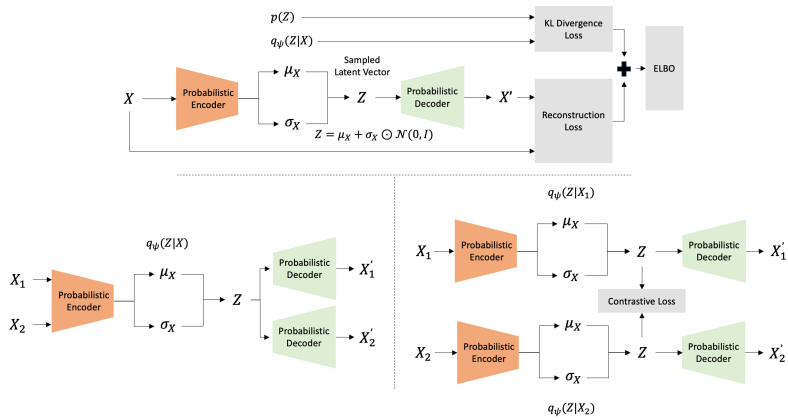

Figure 5.3: A Variational AutoEncoder (VAE) is composed of a probabilistic encoder that maps an input data point to a latent distribution from which a latent vector is sampled, and decoded with a probabilistic decoder with the objective of reconstructing the original input. Top: Standard VAE architecture for unconditional generation. Bottom left: Multimodal VAE architecture with a shared encoder and two unimodal decoders. Bottom right: Contrastive VAE architecture with two unimodal encoders and decoders and a contrastive loss for aligning the latent spaces.

product, and then use it for visual search. The authors trained their image generation network using both a discriminator and a triplet loss to make sure the generated image is not too close to the original query image.

5.3.2 Variational AutoEncoders for Multimodal Recommendation

Proposed by Kingma and Welling (2014), Variational AutoEncoders (VAEs) have been used as a key component of various recommendation systems, including collaborative filtering with implicit feedback (Liang *et al.*, 2018), collaborative filtering with side information (Karamanolakis *et al.*, 2018), and content-based retrieval (Yi *et al.*, 2021). In this subsection, we briefly summarize the formulation of VAEs for unimodal data, and show how it can be extended to multimodal data, including adaptations of VAEs for collaborative filtering and content-based retrieval.

Unimodal VAEs As discussed in Section 3.5.1, VAEs are a class of latent variable models in which a *probabilistic encoder* maps the input

data X to a latent variable Z, and a *probabilistic decoder* maps Z back to X (see Fig. 5.3 top). More specifically, VAEs learn a probability distribution $p(X)$ for data X (e.g., image or text) by positing the existence of a latent variable Z (continuous or discrete) such that $p(X) = \int p_\theta(X \mid Z)p(Z)dZ$. The prior $p(Z)$ is typically assumed to be a standard Gaussian (continuous) or uniform (discrete). The posterior $p_\theta(X \mid Z)$ is typically assumed to be Gaussian or categorical and is implemented by a neural network (encoder) that maps Z to the parameters of $p_\theta(X \mid Z)$, e.g., its mean $\mu_\theta(Z)$. The posterior $p_\theta(Z \mid X)$ is typically intractable and thus approximated by a simpler distribution $q_\psi(Z \mid X)$ (e.g., Gaussian or categorical) implemented by another neural network (encoder), which maps X to, e.g., the mean $\mu_\psi(X)$. The weights of the encoder-decoder pair are learned by maximizing a lower bound for the log likelihood $\log(p(X))$, known as the Evidence Lower Bound (ELBO),

$$\mathcal{L} = \mathbb{E}_{Z \sim q_\psi(Z|X))}\Big[\log p_\theta(X \mid Z) - KL(q_\psi(Z \mid X) \mid\mid p_\theta(Z \mid X))\Big] \quad (5.5)$$

which is the sum of a reconstruction term $\log p_\theta(X \mid Z)$ and a regularization term $KL(q_\psi(Z \mid X) \mid\mid p(Z))$. The variable Z is then used for downstream recommendation tasks.

Multimodal VAEs In the case of multimodal data, say $X = (X^1, X^2)$ consists of both image and text, we can still use the VAE model described so far. However, as we argued in the case of GANs, doing so may require designing a very complex decoder. A better approach is to design multimodal VAEs that leverage unimodal VAEs. For example, as we did in (5.4), we can assume that X^1 and X^2 are conditionally independent given Z, i.e., $p_\theta(X^1, X^2 \mid Z) = p_{\theta_1}(X^1 \mid Z)p_{\theta_2}(X^2 \mid Z)$, so that we can use one decoder per modality. However, since the latent space Z is shared, this requires the design of a shared encoder $q_\psi(Z \mid X^1, X^2)$. Fig. 5.3 (bottom left) shows the design of such a multimodal VAE with a single probabilistic encoder and two modality specific decoders.

To leverage modality specific encoders and decoders pretrained on large datasets, two families of approaches have been proposed. The first family approximates $q_\psi(Z \mid X^1, X^2)$ with a product of experts (Wu and

Goodman, 2018), a mixture of experts (Shi *et al.*, 2019) or a mixture of products of experts (Sutter *et al.*, 2020), allowing one to fuse multiple unimodal encoders into a multimodal one. The second family, partitions the latent space per modality, i.e., $Z = (Z^1, Z^2)$, and assume that $q_\psi(Z \mid X) = q_\psi(Z^1 \mid X^1)q_\psi(Z^2 \mid X^2)$ and $p_\theta(X \mid Z) = p_\theta(X^1 \mid Z^1)p_\theta(X^2 \mid Z^2)$. However, doing so reduces the entire model to two independent VAEs, one per modality, which defeats the purpose of having a multimodal model. ContrastVAE (Wang *et al.*, 2022c) addresses this issue by adding a contrastive loss to the ELBO objective, the InfoNCE loss (Oord *et al.*, 2018), which aligns the latent spaces of the two modalities. Experiments in Wang *et al.* (2022c) show that ContrastVAE improves upon purely contrastive models by adequately modeling data uncertainty and data sparsity, and being robust to perturbations in the latent space.

Multimodal VAEs for collaborative filtering Traditional VAEs for recommendation systems are unimodal in nature as they aim to model user ratings. For example, Liang *et al.* (2018) extends VAEs to collaborative filtering for implicit feedback by using a multinomial likelihood conditional likelihood. However, such models often use the standard Gaussian as a prior, which has been shown to give poor latent representations [16]. Karamanolakis *et al.* (2018) extend VAEs to collaborative filtering with side information. Their key contribution is to replace the standard Gaussian prior in the latent space of the VAE (which is user-agnostic) by a prior that incorporates multimodal user preferences (e.g., user reviews and ratings). The resulting VAE achieves around 30% relative improvement in ranking metric with respect to standard VAEs for collaborative filtering.

Multimodal VAEs for content-based retrieval Yi *et al.* (2021) proposes a multimodal VAE for content-based retrieval. The proposed approach takes three modalities (music, video, and text), maps each modality to a separate latent space using modality-specific encoders, and then aligns these latent spaces via cross-modal generation. More specifically, the video and text modalities are first fused via a product-of-experts model and the fused representation is passed through a cross-modal decoder that generates music. Conversely, the encoding of

music is passed through another crossmodal decoder that generates the visual representation. The resulting representation is trained in 150000 video clips of 3000 different music backgrounds and used to build a music recommendation system.

Graph VAE for Multimodal Recommendation In many applications, multimodal data are better represented by a graph. For example, the graph nodes can be items with hand-crafted or learned features from all modalities, while the graph edges can represent item-item similarities. If we want to learn a VAE for the graph and its features, the encoder needs to be able to process a graph as an input and the decoder needs to generate a graph as an output. Graph neural networks (GNNs) are specialized architectures for processing graphs and can be used as both encoders and decoders. The latent space can be a Gaussian vector, as before, or a graph with one Gaussian vector per node. The resulting model is known as a Graph VAE or GVAE for short (Kipf and Welling, 2016), and has been used in various recommendation systems.

One example is the work if Zhou and Miao (2024), which proposes a Disentangled Graph Variational AutoEncoder (DGVAE) for interpretable multimodal recommendation. DGVAE harnesses contrastive pretraining approaches to map multimodal data to a common space in which user-to-item and user-to-word similarities are used to build an item-to-item graph, which is processed by a GNN. Mutual information maximization is used to regularize the learning objective. Experiments show significant improvements in retrieval performance, especially in terms of the interpretability of the recommendations.

Another example is the work of Chattopadhyay *et al.* (2023), which uses a conditional GVAE to generate decoration recommendations for a room given its type (e.g., bedroom) and its layout (e.g., room elements such as floor and walls). A graph is used to represent both room and furniture layouts, e.g., the nodes capture attributes such as the location, orientation and shape of room and furniture elements, while the edges capture geometric relationships such as relative orientation. Their GVAE then generates a furniture graph, e.g., a collection of furniture items such as bed and night stand that is consistent with the room type and layout, which is then rendered to obtain images of the decorated

room. Experiments on the 3D-FRONT dataset show that their method produces scenes that are diverse and adapted to the room layout.

5.3.3 Diffusion Models for Multimodal Recommendation

Diffusion models (Sohl-Dickstein *et al.*, 2015) have recently emerged as the state-of-the-art approaches for generation of images and other data modalities. They take inspiration from stochastic differential equations (Feller, 1949), dynamical systems (Anderson, 1982) and non-equilibrium thermodynamics (Sohl-Dickstein *et al.*, 2015) to learn highly complex data distributions. Several works have used multimodal diffusion models for recommendation. For example, Li *et al.* (2023e) use diffusion models for sequential recommendation, and Seyfioglu *et al.* (2024) propose a fast diffusion model for virtual try-on that takes a picture with a human model and a white background catalog picture of a garment as input, and places the garment on the model. In this subsection we will discuss the basic diffusion model architecture, its extensions to multimodal data, and applications in content generation.

Unimodal diffusion models The main idea behind a diffusion model is to generate a new image by sampling a random Gaussian vector and transforming it via multiple denoising steps. This is done by defining a forward diffusion process[2] $X \to Z_1 \to Z_2 \to \cdots \to Z_T$ that iteratively maps a data sample X to Gaussian noise Z_T, and a reverse diffusion process $Z_T \to Z_{T-1} \to \cdots \to Z_1 \to X$ that recovers the original data from noise. More specifically, diffusion models are VAEs with a sequential latent space $Z = (Z_1, \ldots, Z_T)$. The VAE encoder $q(Z \mid X) = q(Z_1 \mid X) \prod_{t=2}^{T} q(Z_t \mid Z_{t-1})$ assumes that $Z \mid X$ is Markovian with Gaussian transition probabilities $q(Z_t \mid Z_{t-1})$. The VAE decoder $p(X \mid Z) = p(Z_T)p(X \mid Z_1) \prod_{t=2}^{T} p(Z_{t-1} \mid Z_t)$ assumes that Z is Markovian with Gaussian transition probabilities $p(Z_{t-1} \mid Z_t)$ and $p(Z_T)$ a standard Gaussian. The transition probabilities are parametrized with deep networks whose parameters are learned by maximizing the ELBO objective in (5.5). Once trained, the model can generate high-quality

[2]The forward process is also called corruption or noising process, while the reverse process is also called restoration or denoising process.

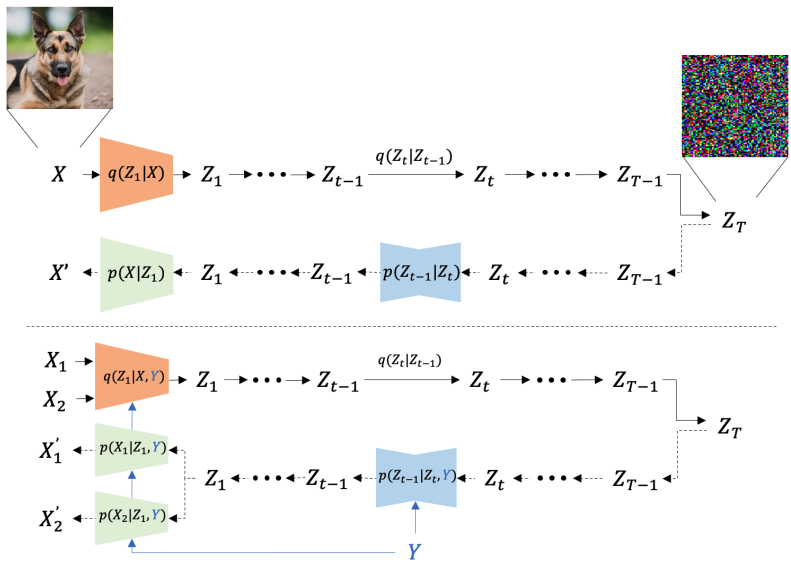

Figure 5.4: A diffusion models consists of a forward process, which iteratively corrupts an input data sample until it becomes Gaussian noise, and a reverse process, which reconstruct the original data from white noise. Top: Latent diffusion model architecture for unconditional generation. The standard diffusion model architecture is obtained by removing the encoder p and the decoder q. Bottom: Conditional multimodal diffusion model with a shared encoder and two unimodal decoders. An unconditional multimodal model is obtained by simply removing the condition on Y.

original data examples by sampling from the noise distribution and simulating the reverse diffusion process.

The original diffusion model for images (Sohl-Dickstein *et al.*, 2015) operates directly in the image space. That is, X is an image and $Z_1, \ldots, Z_T$ are noisy images of the same dimensions as X. Therefore, the encoder does not need to be trained because $q(Z_1 \mid X)$ simply adds noise to the image X. This makes the model simpler, since only the decoder needs to be learned. However, generating high-quality samples requires dividing both the forward and backward processes into small steps, which can be computationally costly when T is large.

To address this issue, stable diffusion (Rombach *et al.*, 2022) uses latent variables $Z_1, \ldots, Z_T$ of smaller dimensions, which makes inference faster, but adds the cost of learning an encoder $q(Z_1 \mid X)$ and decoder

$p(X \mid Z_1)$. In practice, pre-trained models are often used for p and q to avoid the training cost. Figure 5.4 (top) shows the architecture of a diffusion model for image generation. The standard diffusion model operated of Sohl-Dickstein *et al.* (2015) operated in pixel space, and thus did not have the p and q decoder and encoder models.

Even with the addition of latent variables, at inference time diffusion models still require many evaluation steps (Yang *et al.*, 2023b). Recent research has focused on reducing the time spent in the inference process by reducing the number of steps required (Song *et al.*, 2020b; Karras *et al.*, 2022; Dockhorn *et al.*, 2021; Song *et al.*, 2020a; Lu *et al.*, 2022), or training a better sampler to directly select the best possible steps (Watson *et al.*, 2021; Salimans and Ho, 2022; Meng *et al.*, 2023).

Multimodal diffusion models In the case of multimodal data, we could also build a multimodal diffusion model as above, say with $X = (X^1, X^2)$ being images and text. However, two challenges emerge. First, the same challenges of building multimodal encoders and decoders as in VAEs. Second, even if we can build models with separate encoders and decoders per modality, the issue is that diffusion models are not as suitable for text generation as they are for image generation. Specifically, while diffusion models for text generation have been developed, e.g., by using a discrete latent space Z with categorical transition probabilities (Austin *et al.*, 2021), text encoders based on transformers or other sequence-to-sequence models are preferred in practice. As a consequence, multimodal models for both text and images, such as text-to-image generation models, combine text encoders with diffusion models for images. Figure 5.4 (bottom) shows one possible architecture for such a model. As with the other generative approaches, an additional input Y (e.g., a text description) can be used to condition the generation.[3]

Recently, many conditional diffusion models for image generation have been proposed using text and other modalities as the conditioning variables. For example, *DALL-E* (Ramesh *et al.*, 2022; Betker *et al.*, 2023) uses the CLIP (Radford *et al.*, 2021) embedding space as a starting

[3]Note that, unlike the case of VAEs where the conditioning affects both the encoder $q(Z \mid X, Y)$ and the decoder $p(X \mid Z, Y)$, in the case of diffusion models the conditioning does not affect $q(Z_t \mid Z_{t-1})$ because it a simple noising process.

point to generate novel images. To this objective, the authors train a decoder to invert the CLIP representation back to images. Working on a space that jointly represents text and images allows one to apply language-guided image manipulations. Betker *et al.* (2023) improve the quality of the generated images by performing an automated cleaning and improvement of the training image captions with a dedicated captioning model. *Stable Diffusion* (Rombach *et al.*, 2022) is able to generate images from an input text. Since directly training in the pixel space is very computationally demanding, the generative part of Stable Diffusion is trained on a lower-dimension feature space, and relies on a UNet (Ronneberger *et al.*, 2015) autoencoder separately pre-trained on a perceptual loss and a patch-based adversarial objective. To condition the generation based on other modality inputs, such as texts or semantic maps, they train a cross-attention layer to project the new modality inputs to the intermediate layers of the UNet. *Imagen* (Saharia *et al.*, 2022) train a diffusion model for image generation based on a U-Net image model and a T5 text encoder pre-trained only with text. To condition the image generation based on text, the authors find that using cross-attention significantly outperforms other pooling strategies, and achieves high image-text alignment as well as photo-realistic results.

Other works expanded diffusion models in different directions. For example, Zhang *et al.* (2023) increase the controllability of the generated results, Brooks *et al.* (2023) add instruction-following capabilities for image modification, Ruiz *et al.* (2023) improve the consistency of the generated subject's identity by fine-tuning the model with a few images, and Chen *et al.* (2024) propose a multi-modal, multi-task, diffusion model, where multiple input modalities are fused and fed to various decoders to accomplish multiple tasks simultaneously.

Diffusion models have also been used for sound generation (Yang *et al.*, 2023a), video generation (Jeong *et al.*, 2023; Brooks *et al.*, 2024), and other modalities (Kotelnikov *et al.*, 2023; Lin *et al.*, 2023a), or multiple modalities simultaneously (Tang *et al.*, 2024; Ruan *et al.*, 2023). See (Cao *et al.*, 2024) for a recent survey on diffusion models and its applications.

This rapid progress in diffusion model research shows great potential in their usefulness for recommendation applications. Zhu *et al.* (2023)

propose a virtual try-on system based on a diffusion model, which outperforms earlier ones based on GANs. Ma *et al.* (2024) and Jiang *et al.* (2024) use diffusion models to combine multimodal item information with user-item interaction data.

5.3.4 Interactive Multimodal Recommendation Models

As seen in Chapter 4, Large Language Models (LLMs) have been widely used in recommender systems, to do tasks like top-k recommendation, rating prediction or explanation generation (Geng *et al.*, 2022; Wu *et al.*, 2023a; He *et al.*, 2023). In this section we discuss interactive multimodal recommendation models based on LLMs, which have demonstrated impressive generalization capabilities and apparent emergent properties to solve tasks not directly targeted during training (Brown *et al.*, 2020).

Multimodal Large Language Models One approach to designing interactive multimodal recommendation systems is to train or adapt specialized *X-to-text* encoders that allow LLMs to accept multimodal input, such as images (Liu *et al.*, 2024a) or other modalities (Wu *et al.*, 2023b; Tang *et al.*, 2023). These new models are called Multimodal Large Language Models (MLLM), and greatly extend the capabilities of LLMs not only at the input side, by accepting information expressed in different modalities, but also at the output side, where appropriate decoders can be used to allow the model to generate content in various modalities, replacing or complementing the textual answer. Figure 5.5 shows a high-level diagram of an MLLM architecture.

With the addition of multiple modalities, MLLMs can become versatile task solvers for recommendation problems. They provide a natural language interface for users to express their queries in multiple modalities, they can tackle complex zero-shot recommendation tasks thanks to their emergent properties, or orchestrate several sub-systems to obtain the best recommendation, they can also generate fluent natural language explanations for a multi-modal recommendation, or even generate documents in different modalities to help the user visualize the products.

As discussed earlier, given the complexity of training large generative

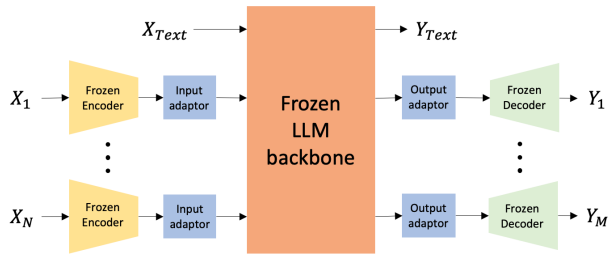

Figure 5.5: High-level architecture of an MLLM model. Each input is processed by a specialized encoder to obtain modality specific features which are then projected to a representation adequate for an LLM backbone via suitable adaptors. Similarly, the output of the LLM is then projected to serve as input for specific generators for each modality via a similar adaptor. The input and output modalities are independent.

models end-to-end, researchers typically assemble systems composed of discriminatively pre-trained components (encoders, decoders and LLM "reasoning" models), usually connected by adaptation layers. These adaptation layers are usually pre-trained using paired data for the different modalities, sometimes together with or followed by some form of parameter-efficient fine-tuning of the base models. For example, Low Rank Adaptation (LoRA) (Hu *et al.*, 2021) freezes the pre-trained parameters and introduces low-rank decomposable trainable adaptation matrices to each transformer layer. This fine-tuning step ensures that representations from different modalities are aligned. In some cases, an existing expert model able to produce textual descriptions of the multi-media content is used in place of adaptation layers (Li *et al.*, 2023c; Yang *et al.*, 2024; Gao *et al.*, 2023; Wang *et al.*, 2024b). This may lead to lower data needs for adaptation, but could also result in information lost in translation.

Controller LLMs Similar to the "dialogue controller" described in Chapter 4.7.2, instead of training adaptation layers, another approach is to allow a "controller" LLM to use external tools (e.g., foundation models, classical recommendation systems, arbitrary functions), to deal with the multi-modal input and output (Yin *et al.*, 2023). This approach has the advantage that it usually involves little or no training. With a carefully constructed prompt explaining all the available tool

capabilities and providing usage examples, the controller LLM can create an execution plan with the multi-modal input and generate the desired result in a zero-shot fashion (Zhang *et al.*, 2024). Once the plan is complete, the controller LLM will have to re-asses the output, and decide if the desired result has been achieved, or further steps are needed. To improve results, researchers have also tried to use instruction-tuning of the controller LLM to improve the tool selection and planning abilities (Yang *et al.*, 2024). An obvious drawback of this approach is that it requires processing multiple rounds of instructions and multiple rounds of (possibly slow) "tool" foundation models to get to the desired result.

Multimodal instruction tuning As seen in Section 4.7.2, instruction tuning is an important step to make LLMs useful task solvers. It requires the creation of datasets of instruction-formatted examples that will be used as training data for the model. These datasets are usually created by extending input-output pairs from multiple multi-modal datasets, such as COCO Captions (Chen *et al.*, 2015), LAION (Schuhmann *et al.*, 2022) or VQAv2 (Goyal *et al.*, 2017), with an instruction text defined using prompt templates for the different tasks (Dai *et al.*, 2024). Tasks can be created, for example, by using annotated bounding boxes to define a spatial relationship question, or using an existing image caption as supervision for an image description request. When no suitable datasets are available, and collecting them from other sources is impractical, researchers use generative models to create the examples through self-instruction (Brooks *et al.*, 2023; Wang *et al.*, 2022b). Models that extend instructions to multiple modalities have recently been proposed (Li *et al.*, 2023d; Wu *et al.*, 2023b). For example, a user could issue a query such as *"modify [image] to convey the feeling of [music]"*. With instruction tuning, MLLMs can be used in dialog systems (c.f. Sec 4.7), enriching the conversation with multimodal understanding and generation.

Any-to-text MLLMs Recently, many MLLMs that add images to the accepted inputs have been developed: Alayrac *et al.* (2022) propose *Flamingo*, which use an LLM and a vision encoder trained with a contrastive loss similar to CLIP to build a model able to meet the

state of the art performance in various image and video tasks. Liu *et al.* (2024a) propose *Llava*, an instruction-tuned multi-modal LLM that is able to accept input in both text and image format, and produce useful textual responses. The authors connect the grid visual embeddings from the last layers of the CLIP image encoder (Radford *et al.*, 2021) with the Vicuna language decoder (Chiang *et al.*, 2023) using a simple linear adaptation layer, and fine-tunning the model end to end (they keep the visual encoder weights frozen). To train the model, instruction-following training data is generated using GPT-4 (Achiam *et al.*, 2023) and image-caption datasets, such as COCO (Chen *et al.*, 2015). Liu *et al.* (2023) change the connection layer from a linear projection to a two-layer MLP and obtain better results. Li *et al.* (2023b) propose *BLIP-2*, introducing a lightweight Query-Transformer (Q-Former), which consists of two transformer modules, to bridge the modality gap between the image encoder output and the LLM, and allow prompts to include both text and image.

Although image is the modality that received the most attention, some works have addressed adding audio (Deshmukh *et al.*, 2023; Kong *et al.*, 2024), and multiple modalities, like text, image, audio or video (Han *et al.*, 2023; Moon *et al.*, 2023; Lyu *et al.*, 2023). However, these models are still limited to generating only text output.

Any-to-any MLLMs As mentioned earlier, to overcome the single-modality output limitation, authors have proposed systems that can both absorb information in multiple modalities, as well as generate response content in different modalities. For example, Next-GPT (Wu *et al.*, 2023b) attempts any-to-any modality conversion through an MM-LLM by using state-of-the-art encoders and decoders, connected to the LLM by thin adapter layers. Multi-modality switching instruction tunning is learned using a custom dataset of 5000 high quality samples. After warming up the adaptation layers, the whole system is trained using LoRA with the modality-switching dataset. For input, the authors use ImageBind (Girdhar *et al.*, 2023), which has been trained to produce aligned representations for image, audio and video, among other modalities, and adapt it using a linear layer to the Vicuna LLM, that does the core reasoning/instruction-following. For the output, small

modality-specific transformers are trained to produce the input for three state-of-the-art decoder models, for audio, image and video. Tang *et al.* (2023) use a similar approach, with Llama2 (Touvron *et al.*, 2023) as the core LLM and state-of-the-art diffusion models to generate the multi-modal outputs.

Several companies have released proprietary generalist MLLM-powered chatbots, like OpenAI GPT-4 (Achiam *et al.*, 2023) and Google Gemini (Team *et al.*, 2023) or Anthropic Claude (Team, 2024). Even though these models are not explicitly trained as recommender systems, they are able to produce a variety of recommendation results, including shopping recommendations. For example, Gemini can receive images, audio and video as input, recognize objects in them, provide general advice for product understanding, and recommend products based on customer input.

5.4 Applications of Multimodal Recommendation Systems

Recent developments in multi-modal generative models open the door to many applications in recommender systems. In this section we review some of the most promising directions, in areas including e-commerce, in-context product visualization, marketing, online streaming, and travel and service recommendations.

E-commerce One of the most direct applications of generative multi-modal models for recommendation is e-commerce, where there is a large volume of product and customer data available that can be used to benefit the customer recommendations. Applications range from improving product images (Corneanu *et al.*, 2024), names and descriptions (Novgorodov *et al.*, 2019; Shao *et al.*, 2021), to generating reviews Truong and Lauw (2019) and review summaries (Schermerhorn, 2023), learning to generate better recommendation (Xiao *et al.*, 2022; Liu *et al.*, 2024b; Karra and Tulabandhula, 2024), and answering user questions (Deng *et al.*, 2022).

Karra and Tulabandhula (2024) propose to use multimodal large language models to improve recommendations by better understanding the behavior of users in e-commerce websites. As the user navigates

during a browsing session, high-frequency screenshots are captured and provided to an MLLM together with specific prompts requesting to extract information such as price ranges, product categories and brand preferences, to generate a user behavioral summary. Next, this summary is provided to an LLM with tool-using abilities to derive features and constraints from the input, and use a recommender system to generate the final recommendation. Liu *et al.* (2024b) describe the limitations of current MLLM when used with multiple images as input in the prompt. To improve the performance, they propose to process the list of products interacted by the user as pairs of image and title to obtain text descriptions of the products. These descriptions can then be used in lieu of the images when using the interaction history as in-context-learning to generate new recommendations for a user with an MLLM. Truong and Lauw (2019) propose a system to generate multi-modal reviews. The proposed system uses item and user embeddings, obtained via matrix factorization, to predict the rating and compose a review text with a Long Short-Term Memory network. If a review image is available, it is also used to condition the text generation.

In-context product visualization Applications such as "virtual try on" or "view in your room" augment an image or video with products such as clothes (Yuan *et al.*, 2013; Han *et al.*, 2018), sunglasses, or even makeup (Borges and Morimoto, 2019; Prinzivalli, 2019) to help users visualize how they would look in themselves, or how furniture or appliances would look in the context of their home (Reuksupasompon *et al.*, 2018; Perez, 2020; Berthiaume, 2023), before making a purchase decision. A traditional approach for these tasks is Augmented Reality (AR), that mixes real images, obtained from a camera feed, with virtual objects to generate novel views in real time. While AR has been used in numerous applications, ranging from education (Billinghurst, 2002) to assisting surgeons in medical operations (Dennler *et al.*, 2021), recent diffusion-based image generation models can be used to further improve virtual-try-on experiences (Wang *et al.*, 2024a; Xu *et al.*, 2024b; Wang and Ye, 2024), or generate outfits to try out (Xu *et al.*, 2024a), and make them more controllable (Seyfioglu *et al.*, 2024).

Marketing In marketing, multimodal generative models can be used to create personalized advertisement images and videos from product imagery and customer preferences to increase the probability of engagement (Wang *et al.*, 2023; Chen *et al.*, 2021a). Wei *et al.* (2022) generates personalized bundles and creates a customized image for display. Shilova *et al.* (2023) fine-tune a Stable Diffusion model to generate personalized images by outpainting input images without modifying the targeted object. For training, they leverage a U^2-Net segmentation network, and a BLIP model to generate masks and captions for a collection of training images, that the model will then learn to reconstruct. With adequate guardrails in place, generative models could also be used to synthesize personalized multimodal ad content like text (Loukili *et al.*, 2023), images (Mayahi and Vidrih, 2022) or video (Liu and Yu, 2023).

Streaming services Online video and audio streaming services strive to recommend the most valuable multimedia content to each user in order to maximize usage, ad revenue or click-through rate. Long and short-form video, music, audiobooks, podcasts and radio have different recommendation requirements, but the very content to recommend comes in multiple modalities that can be used to improve the suggestions. Even though most works on streaming content recommendation rely on user behavior and content metadata, recent works have applied multimodal learning to audio (Jones, 2023; Chen *et al.*, 2021b; Huang *et al.*, 2020; Deldjoo *et al.*, 2024) and video (Lei *et al.*, 2021; Wei *et al.*, 2019; Yi *et al.*, 2022; Sun *et al.*, 2022) recommendation. Due to their extensive pre-training, large multi-modal generative models can further enhance the user experience in a content streaming recommendation setting by blending content understanding with personalization and generation, allowing them to complete tasks like answering to fine-grained content-related questions in natural language (e.g., "Does this movie contain a car chase scene that I will like?"), or generating personalized content to fulfill a user request (e.g., audio and music generation (Briot *et al.*, 2017; Lam *et al.*, 2024; Vyas *et al.*, 2023; Dhariwal *et al.*, 2020)).

Travel and service recommendations Services ranging from theme parks and concert venues to restaurants, auto mechanics, and laundry

services receive customer ratings, reviews and clicks in many online platforms. Better understanding of contextual details such as location characteristics, services offered, past user experiences and popularity factors through multi-modal information, could lead to better and more personalized recommendations. Furthermore, future work could tackle comprehensive products such as interactive recommendation systems able to help users through the whole process of planning and booking complex multi-faceted events such as weddings (e.g., venue, menu, decoration, music) or travel (e.g., destination, transport, hotel, restaurant, activities, practical tips) in a conversational way (Xie *et al.*, 2024), accepting and incorporating user feedback and explaining the recommendations. Yan *et al.* (2023) generates explanations for recommendations focusing on making them informative and diverse. For that, they start by selecting a set of images for a given user and business using a detrimental point process that leverages CLIP features from the user history, and the business images. Then, they use a GPT-2-powered multi-modal decoder, trained with a personalized cross-modal contrastive loss, to generate natural language explanations. The results show that the proposed method produces more informative and diverse explanations compared to text-only alternatives.

References

Achiam, J., S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.* (2023). "Gpt-4 technical report". *arXiv preprint arXiv:2303.08774*.

Alayrac, J.-B., J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, *et al.* (2022). "Flamingo: a visual language model for few-shot learning". *Advances in neural information processing systems*. 35: 23716–23736.

Alpay, T., S. Magg, P. Broze, and D. Speck. (2023). "Multimodal video retrieval with CLIP: a user study". *Information Retrieval Journal*. 26(1-2).

Anderson, B. D. (1982). "Reverse-time diffusion equation models". *Stochastic Processes and their Applications*. 12(3): 313–326.

Arjovsky, M., S. Chintala, and L. Bottou. (2017). "Wasserstein generative adversarial networks". In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70. ICML'17*. Sydney, NSW, Australia: JMLR.org. 214–223.

Austin, J., D. D. Johnson, J. Ho, D. Tarlow, and R. van den Berg. (2021). "Structured Denoising Diffusion Models in Discrete State-Spaces". *CoRR*. abs/2107.03006. arXiv: 2107.03006. URL: https://arxiv.org/abs/2107.03006.

Baldrati, A., L. Agnolucci, M. Bertini, and A. Del Bimbo. (2023). "Zero-shot composed image retrieval with textual inversion". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15338–15347.

Bansal, H., N. Singhi, Y. Yang, F. Yin, A. Grover, and K.-W. Chang. (2023). "Cleanclip: Mitigating data poisoning attacks in multimodal contrastive learning". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 112–123.

Berthiaume, D. (2023). "Amazon enhances online shopping experience with AI, AR". https://chainstoreage.com/amazon-enhances-online-shopping-experience-ai-ar.

Betker, J., G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, Y. Guo, *et al.* (2023). "Improving image generation with better captions". *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf.* 2(3): 8.

Billinghurst, M. (2002). "Augmented reality in education". *New horizons for learning*. 12(5): 1–5.

Borges, A. d. F. S. and C. H. Morimoto. (2019). "A virtual makeup augmented reality system". In: *2019 21st Symposium on Virtual and Augmented Reality (SVR)*. IEEE. 34–42.

Briot, J.-P., G. Hadjeres, and F.-D. Pachet. (2017). "Deep learning techniques for music generation–a survey". *arXiv preprint arXiv:1709.01620*.

Brooks, T., A. Holynski, and A. A. Efros. (2023). "Instructpix2pix: Learning to follow image editing instructions". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18392–18402.

Brooks, T., B. Peebles, C. Holmes, W. DePue, Y. Guo, L. Jing, D. Schnurr, J. Taylor, T. Luhman, E. Luhman, C. Ng, R. Wang, and A. Ramesh. (2024). "Video generation models as world simulators". URL: https://openai.com/research/video-generation-models-as-world-simulators.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.* (2020). "Language models are few-shot learners". *Advances in neural information processing systems*. 33: 1877–1901.

Cao, H., C. Tan, Z. Gao, Y. Xu, G. Chen, P.-A. Heng, and S. Z. Li. (2024). "A survey on generative diffusion models". *IEEE Transactions on Knowledge and Data Engineering*.

Cao, L., B. Zhang, C. Chen, Y. Yang, X. Du, W. Zhang, Z. Lu, and Y. Zheng. (2023). "Less is More: Removing Text-regions Improves CLIP Training Efficiency and Robustness". *arXiv preprint arXiv:2305.05095*.

Cao, Z., G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. (2019). "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields". *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Carlsson, F., P. Eisen, F. Rekathati, and M. Sahlgren. (2022). "Cross-lingual and multilingual clip". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. 6848–6854.

Caron, M., H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. (2021). "Emerging Properties in Self-Supervised Vision Transformers". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 9650–9660.

Chakraborty, T., U. R. K. S, S. M. Naik, M. Panja, and B. Manvitha. (2024). "Ten years of generative adversarial nets (GANs): a survey of the state-of-the-art". *Machine Learning: Science and Technology*. 5(1): 011001. DOI: 10.1088/2632-2153/ad1f77. URL: https://dx.doi.org/10.1088/2632-2153/ad1f77.

Chattopadhyay, A., X. Zhang, D. P. Wipf, H. Arora, and R. Vidal. (2023). "Learning graph variational autoencoders with constraints and structured priors for conditional indoor 3d scene generation". In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 785–794.

Chen, C., H. Ding, B. Sisman, Y. Xu, O. Xie, B. Yao, S. Tran, and B. Zeng. (2024). "Diffusion models for multi-modal generative modeling". In: *ICLR 2024*. URL: https://www.amazon.science/publications/diffusion-models-for-multi-modal-generative-modeling.

Chen, C., B. Zhang, L. Cao, J. Shen, T. Gunter, A. M. Jose, A. Toshev, J. Shlens, R. Pang, and Y. Yang. (2023a). "STAIR: Learning Sparse Text and Image Representation in Grounded Tokens". *arXiv preprint arXiv:2301.13081*.

Chen, Z.-j., W. Chen, J. Xu, Z. Liu, and W. Zhang. (2023b). "Beyond Semantics: Learning a Behavior Augmented Relevance Model with Self-supervised Learning". *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. URL: https://api.semanticscholar.org/CorpusID:260775754.

Chen, J., J. Xu, G. Jiang, T. Ge, Z. Zhang, D. Lian, and K. Zheng. (2021a). "Automated creative optimization for e-commerce advertising". In: *Proceedings of the Web Conference 2021*. 2304–2313.

Chen, K., B. Liang, X. Ma, and M. Gu. (2021b). "Learning audio embeddings with user listening data for content-based music recommendation". In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 3015–3019.

Chen, M., A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever. (2020a). "Generative pretraining from pixels". In: *International conference on machine learning*. PMLR. 1691–1703.

Chen, T., S. Kornblith, M. Norouzi, and G. Hinton. (2020b). "A simple framework for contrastive learning of visual representations". In: *International conference on machine learning*. PMLR. 1597–1607.

Chen, X., Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. (2016). "InfoGAN: interpretable representation learning by information maximizing generative adversarial nets". In: *Proceedings of the 30th International Conference on Neural Information Processing Systems. NIPS'16*. Barcelona, Spain: Curran Associates Inc. 2180–2188. ISBN: 9781510838819.

Chen, X., H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. (2015). "Microsoft coco captions: Data collection and evaluation server". *arXiv preprint arXiv:1504.00325*.

Cheng, Y., R. Wang, Z. Pan, R. Feng, and Y. Zhang. (2020). "Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning". In: *Proceedings of the 28th ACM International Conference on Multimedia*. 3884–3892.

Cherti, M., R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev. (2023). "Reproducible scaling laws for contrastive language-image learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2818–2829.

Chiang, W.-L., Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing. (2023). "Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality". URL: https://lmsys.org/blog/2023-03-30-vicuna/.

Corneanu, C., R. Gadde, and A. M. Martinez. (2024). "LatentPaint: Image Inpainting in Latent Space With Diffusion Models". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 4334–4343.

Croitoru, F.-A., V. Hondru, R. T. Ionescu, and M. Shah. (2023). "Diffusion models in vision: A survey". *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Dai, W., J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. N. Fung, and S. Hoi. (2024). "Instructblip: Towards general-purpose vision-language models with instruction tuning". *Advances in Neural Information Processing Systems*. 36.

Dehdashtian, S., L. Wang, and V. N. Boddeti. (2024). "FAIRERCLIP: DEBIASING ZERO-SHOT PREDICTIONS OF CLIP IN RKHSS". In: *ICLR2024? VERIFY*.

Deldjoo, Y., M. Schedl, and P. Knees. (2024). "Content-driven music recommendation: Evolution, state of the art, and challenges". *Computer Science Review*. 51: 100618.

Deng, Y., Y. Li, W. Zhang, B. Ding, and W. Lam. (2022). "Toward personalized answer generation in e-commerce via multi-perspective preference modeling". *ACM Transactions on Information Systems (TOIS)*. 40(4): 1–28.

Dennler, C., D. E. Bauer, A.-G. Scheibler, J. Spirig, T. Götschi, P. Fürnstahl, and M. Farshad. (2021). "Augmented reality in the operating room: a clinical feasibility study". *BMC musculoskeletal disorders*. 22(1): 451.

Deshmukh, S., B. Elizalde, R. Singh, and H. Wang. (2023). "Pengi: An audio language model for audio tasks". *Advances in Neural Information Processing Systems*. 36: 18090–18108.

Dhariwal, P., H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever. (2020). "Jukebox: A generative model for music". *arXiv preprint arXiv:2005.00341*.

Dockhorn, T., A. Vahdat, and K. Kreis. (2021). "Score-based generative modeling with critically-damped langevin diffusion". *arXiv preprint arXiv:2112.07068*.

Du, M., A. Ramisa, A. K. KC, S. Chanda, M. Wang, N. Rajesh, S. Li, Y. Hu, T. Zhou, N. Lakshminarayana, *et al.* (2022). "Amazon shop the look: A visual search system for fashion and home". In: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2822–2830.

Fan, L., D. Krishnan, P. Isola, D. Katabi, and Y. Tian. (2023). "Improving CLIP Training with Language Rewrites". *arXiv preprint arXiv:2305.20088*.

Feller, W. (1949). "On the theory of stochastic processes, with particular reference to applications". In: *Berkeley Symposium on Mathematical Statistics and Probability*. 403–432.

Gao, M., J. Zhang, J. Yu, J. Li, J. Wen, and Q. Xiong. (2021). "Recommender systems based on generative adversarial networks: A problem-driven perspective". *Information Sciences*. 546: 1166–1185.

Gao, P., J. Han, R. Zhang, Z. Lin, S. Geng, A. Zhou, W. Zhang, P. Lu, C. He, X. Yue, *et al.* (2023). "Llama-adapter v2: Parameter-efficient visual instruction model". *arXiv preprint arXiv:2304.15010*.

Geng, S., S. Liu, Z. Fu, Y. Ge, and Y. Zhang. (2022). "Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5)". In: *Proceedings of the 16th ACM Conference on Recommender Systems*. 299–315.

Girdhar, R., A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra. (2023). "Imagebind: One embedding space to bind them all". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15180–15190.

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. (2014). "Generative adversarial nets". In: *Advances in Neural Information Processing Systems*.

Goyal, Y., T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. (2017). "Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Grill, J.-B., F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko. (2020). "Bootstrap your own latent a new approach to self-supervised learning". In: *Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS'20*. Vancouver, BC, Canada: Curran Associates Inc. ISBN: 9781713829546.

Gu, X., T.-Y. Lin, W. Kuo, and Y. Cui. (2021). "Open-vocabulary Object Detection via Vision and Language Knowledge Distillation". In: *International Conference on Learning Representations*. URL: https://api.semanticscholar.org/CorpusID:238744187.

Guo, W., J. Wang, and S. Wang. (2019). "Deep multimodal representation learning: A survey". *Ieee Access*. 7: 63373–63394.

Gutmann, M. and A. Hyvärinen. (2010). "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models". In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings. 297–304.

Hager, P., M. J. Menten, and D. Rueckert. (2023). "Best of Both Worlds: Multimodal Contrastive Learning with Tabular and Imaging Data". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 23924–23935.

Han, J., R. Zhang, W. Shao, P. Gao, P. Xu, H. Xiao, K. Zhang, C. Liu, S. Wen, Z. Guo, *et al.* (2023). "Imagebind-llm: Multi-modality instruction tuning". *arXiv preprint arXiv:2309.03905*.

Han, X., Z. Wu, Z. Wu, R. Yu, and L. S. Davis. (2018). "Viton: An image-based virtual try-on network". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7543–7552.

He, K., X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. (2022). "Masked autoencoders are scalable vision learners". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16000–16009.

He, K., H. Fan, Y. Wu, S. Xie, and R. Girshick. (2020). "Momentum contrast for unsupervised visual representation learning". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9729–9738.

He, Z., Z. Xie, R. Jha, H. Steck, D. Liang, Y. Feng, B. P. Majumder, N. Kallus, and J. McAuley. (2023). "Large language models as zero-shot conversational recommenders". *arXiv preprint arXiv:2308.10053*.

Hendriksen, M., M. Bleeker, S. Vakulenko, N. van Noord, E. Kuiper, and M. de Rijke. (2022). "Extending CLIP for Category-to-image Retrieval in E-commerce". In: *European Conference on Information Retrieval*. Springer. 289–303.

Hu, E. J., Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. (2021). "Lora: Low-rank adaptation of large language models". *arXiv preprint arXiv:2106.09685*.

Huang, J.-T., C.-L. Lu, P.-K. Chang, C.-I. Huang, C.-C. Hsu, P.-J. Huang, H.-C. Wang, *et al.* (2021). "Cross-modal contrastive learning of representations for navigation using lightweight, low-cost millimeter wave radar for adverse environmental conditions". *IEEE Robotics and Automation Letters*. 6(2): 3333–3340.

Huang, Q., A. Jansen, L. Zhang, D. P. Ellis, R. A. Saurous, and J. Anderson. (2020). "Large-scale weakly-supervised content embeddings for music recommendation and tagging". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 8364–8368.

Huang, W. (2023). "Multimodal Contrastive Learning and Tabular Attention for Automated Alzheimer's Disease Prediction". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2473–2482.

Huang, X., H. Zhou, K. Yao, and K. Han. (2024). "FROSTER: Frozen CLIP is A Strong Teacher for Open-Vocabulary Action Recognition". In: *The Twelfth International Conference on Learning Representations*.

Huang, X., A. Mallya, T.-C. Wang, and M.-Y. Liu. (2022). "Multi-modal conditional image synthesis with product-of-experts gans". In: *European Conference on Computer Vision*. Springer. 91–109.

Isola, P., J.-Y. Zhu, T. Zhou, and A. A. Efros. (2017). "Image-to-Image Translation with Conditional Adversarial Networks". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5967–5976. DOI: 10.1109/CVPR.2017.632.

Jang, E., S. Gu, and B. Poole. (2016). "Categorical Reparameterization with Gumbel-Softmax". In: *International Conference on Learning Representations*.

Jeong, Y., W. Ryoo, S. Lee, D. Seo, W. Byeon, S. Kim, and J. Kim. (2023). "The power of sound (tpos): Audio reactive video generation with stable diffusion". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7822–7832.

Jia, C., Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig. (2021). "Scaling up visual and vision-language representation learning with noisy text supervision". In: *International conference on machine learning*. PMLR. 4904–4916.

Jiang, Y., L. Xia, W. Wei, D. Luo, K. Lin, and C. Huang. (2024). "DiffMM: Multi-Modal Diffusion Model for Recommendation". *arXiv preprint arXiv:2406.11781*.

Jones, R. (2023). "Learning to Understand Audio and Multimodal Content". In: *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining. WSDM '23.* , Singapore, Singapore, Association for Computing Machinery. 4–5. ISBN: 9781450394079. DOI: 10.1145/3539597.3572333. URL: https://doi.org/10.1145/3539597.3572333.

Karamanolakis, G., K. R. Cherian, A. R. Narayan, J. Yuan, D. Tang, and T. Jebara. (2018). "Item recommendation with variational autoencoders and heterogeneous priors". In: *Proceedings of the 3rd Workshop on Deep Learning for Recommender Systems*. 10–14.

Karra, S. R. and T. Tulabandhula. (2024). "InteraRec: Interactive Recommendations Using Multimodal Large Language Models". *arXiv preprint arXiv:2403.00822*.

Karras, T., M. Aittala, T. Aila, and S. Laine. (2022). "Elucidating the design space of diffusion-based generative models". *Advances in Neural Information Processing Systems*. 35: 26565–26577.

Kingma, D. P. and M. Welling. (2014). "Auto-Encoding Variational Bayes". In: *Proceedings of the 2nd International Conference on Learning Representations, ICLR 2014*.

Kipf, T. N. and M. Welling. (2016). "Variational graph auto-encoders". *arXiv preprint arXiv:1611.07308*.

Kirillov, A., E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.* (2023). "Segment anything". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4015–4026.

Kong, Z., A. Goel, R. Badlani, W. Ping, R. Valle, and B. Catanzaro. (2024). "Audio Flamingo: A Novel Audio Language Model with Few-Shot Learning and Dialogue Abilities". *arXiv preprint arXiv:2402.01831*.

Kotelnikov, A., D. Baranchuk, I. Rubachev, and A. Babenko. (2023). "TabDDPM: Modelling Tabular Data with Diffusion Models". In: *Proceedings of the 40th International Conference on Machine Learning*. Ed. by A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett. Vol. 202. *Proceedings of Machine Learning Research*. PMLR. 17564–17579. URL: https://proceedings.mlr.press/v202/kotelnikov23a.html.

Lam, M. W., Q. Tian, T. Li, Z. Yin, S. Feng, M. Tu, Y. Ji, R. Xia, M. Ma, X. Song, *et al.* (2024). "Efficient neural music generation". *Advances in Neural Information Processing Systems*. 36.

Lei, C., Y. Liu, L. Zhang, G. Wang, H. Tang, H. Li, and C. Miao. (2021). "SEMI: A Sequential Multi-Modal Information Transfer Network for E-Commerce Micro-Video Recommendations". In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. KDD '21*. Virtual Event, Singapore: Association for Computing Machinery. 3161–3171. ISBN: 9781450383325. DOI: 10.1145/3447548.3467189. URL: https://doi.org/10.1145/3447548.3467189.

Li, C., Z. Gan, Z. Yang, J. Yang, L. Li, L. Wang, and J. Gao. (2023a). "Multimodal foundation models: From specialists to general-purpose assistants". *arXiv preprint arXiv:2309.10020*. 1(2): 2.

Li, J., D. Li, S. Savarese, and S. Hoi. (2023b). "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models". In: *International conference on machine learning*. PMLR. 19730–19742.

Li, J., D. Li, C. Xiong, and S. Hoi. (2022). "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation". In: *International Conference on Machine Learning*. PMLR. 12888–12900.

Li, J., R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi. (2021). "Align before fuse: Vision and language representation learning with momentum distillation". *Advances in neural information processing systems*. 34: 9694–9705.

Li, K., Y. He, Y. Wang, Y. Li, W. Wang, P. Luo, Y. Wang, L. Wang, and Y. Qiao. (2023c). "Videochat: Chat-centric video understanding". *arXiv preprint arXiv:2305.06355*.

Li, S., H. Singh, and A. Grover. (2023d). "InstructAny2Pix: Flexible Visual Editing via Multimodal Instruction Following". *arXiv preprint arXiv:2312.06738*.

Li, Z., A. Sun, and C. Li. (2023e). "DiffuRec: A Diffusion Model for Sequential Recommendation". *ACM Transactions on Information Systems*. 42(Dec.): 1–28. DOI: 10.1145/3631116.

Liang, D., R. G. Krishnan, M. D. Hoffman, and T. Jebara. (2018). "Variational autoencoders for collaborative filtering". In: *Proceedings of the 2018 world wide web conference*. 689–698.

Lin, L., Z. Li, R. Li, X. Li, and J. Gao. (2023a). "Diffusion models for time-series applications: a survey". *Frontiers of Information Technology & Electronic Engineering*: 1–23.

Lin, Z., Y. Tan, Y. Zhan, W. Liu, F. Wang, C. Chen, S. Wang, and C. Yang. (2023b). "Contrastive Intra-and Inter-Modality Generation for Enhancing Incomplete Multimedia Recommendation". In: *Proceedings of the 31st ACM International Conference on Multimedia*. 6234–6242.

Liu, C. and H. Yu. (2023). "Ai-empowered persuasive video generation: A survey". *ACM Computing Surveys*. 55(13s): 1–31.

Liu, H., C. Li, Y. Li, and Y. J. Lee. (2023). "Improved baselines with visual instruction tuning". *arXiv preprint arXiv:2310.03744*.

Liu, H., C. Li, Q. Wu, and Y. J. Lee. (2024a). "Visual instruction tuning". *Advances in neural information processing systems*. 36.

Liu, L., H. Zhang, and D. Zhou. (2021). "Clothing generation by multi-modal embedding: A compatibility matrix-regularized GAN model". *Image and Vision Computing*. 107: 104097.

Liu, Y., W. Chen, L. Liu, and M. S. Lew. (2019). "SwapGAN: A multistage generative approach for person-to-person fashion style transfer". *IEEE Transactions on Multimedia*. 21(9): 2209–2222.

Liu, Y., Y. Wang, L. Sun, and P. S. Yu. (2024b). "Rec-GPT4V: Multimodal Recommendation with Large Vision-Language Models". *arXiv preprint arXiv:2402.08670*.

Loukili, S., A. Fennan, and L. Elaachak. (2023). "Applications of Text Generation in Digital Marketing: a review". In: *Proceedings of the 6th International Conference on Networking, Intelligent Systems & Security. NISS '23.* , Larache, Morocco, Association for Computing Machinery. ISBN: 9798400700194. DOI: 10.1145/3607720.3608451. URL: https://doi.org/10.1145/3607720.3608451.

Lu, C., Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu. (2022). "Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps". *Advances in Neural Information Processing Systems*. 35: 5775–5787.

Lyu, C., M. Wu, L. Wang, X. Huang, B. Liu, Z. Du, S. Shi, and Z. Tu. (2023). "Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration". *arXiv preprint arXiv:2306.09093*.

Ma, H., Y. Yang, L. Meng, R. Xie, and X. Meng. (2024). "Multimodal Conditioned Diffusion Model for Recommendation". In: *Companion Proceedings of the ACM on Web Conference 2024. WWW '24.* , Singapore, Singapore, Association for Computing Machinery. 1733–1740. ISBN: 9798400701726. DOI: 10.1145/3589335.3651956. URL: https://doi.org/10.1145/3589335.3651956.

Mayahi, S. and M. Vidrih. (2022). "The impact of generative ai on the future of visual content marketing". *arXiv preprint arXiv:2211.12660*.

Mehta, R. (2024). "Amazon announces Rufus, a new generative AI-powered conversational shopping experience". https://www.aboutamazon.com/news/retail/amazon-rufus.

Meng, C., R. Rombach, R. Gao, D. Kingma, S. Ermon, J. Ho, and T. Salimans. (2023). "On distillation of guided diffusion models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14297–14306.

Mirza, M. and S. Osindero. (2014). "Conditional Generative Adversarial Nets". *CoRR*. abs/1411.1784. arXiv: 1411.1784. URL: http://arxiv.org/abs/1411.1784.

Moon, S., A. Madotto, Z. Lin, T. Nagarajan, M. Smith, S. Jain, C.-F. Yeh, P. Murugesan, P. Heidari, Y. Liu, *et al.* (2023). "Anymal: An efficient and scalable any-modality augmented language model". *arXiv preprint arXiv:2309.16058*.

Nguyen, Q., T. Vu, A. Tran, and K. Nguyen. (2024). "Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation". *Advances in Neural Information Processing Systems*. 36.

Novack, Z., J. McAuley, Z. C. Lipton, and S. Garg. (2023). "Chils: Zero-shot image classification with hierarchical label sets". In: *International Conference on Machine Learning*. PMLR. 26342–26362.

Novgorodov, S., I. Guy, G. Elad, and K. Radinsky. (2019). "Generating product descriptions from user reviews". In: *The world wide web conference*. 1354–1364.

Oord, A. v. d., Y. Li, and O. Vinyals. (2018). "Representation learning with contrastive predictive coding". *arXiv preprint arXiv:1807.03748*.

Pandey, N. and A. Savakis. (2020). "Poly-GAN: Multi-conditioned GAN for fashion synthesis". *Neurocomputing*. 414: 356–364.

Perez, S. (2020). "Amazon rolls out a new AR shopping feature for viewing multiple items at once". https://techcrunch.com/2020/08/25/amazon-rolls-out-a-new-ar-shopping-feature-for-viewing-multiple-items-at-once/.

Pham, H., Z. Dai, G. Ghiasi, K. Kawaguchi, H. Liu, A. W. Yu, J. Yu, Y.-T. Chen, M.-T. Luong, Y. Wu, *et al.* (2023). "Combined scaling for zero-shot transfer learning". *Neurocomputing*. 555: 126658.

Prinzivalli, L. (2019). “L’Oréal and Amazon Just Rolled Out a Virtual Makeup Try-on Functionality”. https://www.allure.com/story/loreal-amazon-modiface-virtual-try-on-lipstick.

Radford, A., J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.* (2021). “Learning transferable visual models from natural language supervision”. In: *International conference on machine learning*. PMLR. 8748–8763.

Rahate, A., R. Walambe, S. Ramanna, and K. Kotecha. (2022). “Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions”. *Information Fusion*. 81: 203–239.

Ramesh, A., P. Dhariwal, A. Nichol, C. Chu, and M. Chen. (2022). “Hierarchical text-conditional image generation with clip latents”. *arXiv preprint arXiv:2204.06125*. 1(2): 3.

Reuksupasompon, P., M. Aruncharathorn, and S. Vittayakorn. (2018). “Ar development for room design”. In: *2018 15th International Joint Conference on Computer Science and Software Engineering (JCSSE)*. IEEE. 1–6.

Rombach, R., A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. (2022). “High-resolution image synthesis with latent diffusion models”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

Ronneberger, O., P. Fischer, and T. Brox. (2015). “U-net: Convolutional networks for biomedical image segmentation”. In: *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer. 234–241.

Rosenbaum, A., S. Soltan, W. Hamza, A. Saffari, M. Damonte, and I. Groves. (2022). “CLASP: Few-Shot Cross-Lingual Data Augmentation for Semantic Parsing”. *AACL-IJCNLP 2022*: 444.

Ruan, L., Y. Ma, H. Yang, H. He, B. Liu, J. Fu, N. J. Yuan, Q. Jin, and B. Guo. (2023). “Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10219–10228.

Ruiz, N., Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman. (2023). "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22500–22510.

Saeed, A., D. Grangier, and N. Zeghidour. (2021). "Contrastive learning of general-purpose audio representations". In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 3875–3879.

Saharia, C., W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, *et al.* (2022). "Photorealistic text-to-image diffusion models with deep language understanding". *Advances in Neural Information Processing Systems*. 35: 36479–36494.

Salimans, T. and J. Ho. (2022). "Progressive distillation for fast sampling of diffusion models". *arXiv preprint arXiv:2202.00512*.

Schermerhorn, V. (2023). "How Amazon continues to improve the customer reviews experience with generative AI". https://www.aboutamazon.com/news/amazon-ai/amazon-improves-customer-reviews-with-generative-ai.

Schuhmann, C., R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, *et al.* (2022). "Laion-5b: An open large-scale dataset for training next generation image-text models". *Advances in Neural Information Processing Systems*. 35: 25278–25294.

Sevegnani, K., A. Seshadri, T. Wang, A. Beniwal, J. McAuley, A. Lu, and G. Medioni. (2022). "Contrastive learning for interactive recommendation in fashion". *arXiv preprint arXiv:2207.12033*.

Seyfioglu, M. S., K. Bouyarmane, S. Kumar, A. Tavanaei, and I. B. Tutar. (2024). "Diffuse to Choose: Enriching Image Conditioned Inpainting in Latent Diffusion Models for Virtual Try-All". arXiv: 2401.13795 [cs.CV].

Shao, H., J. Wang, H. Lin, X. Zhang, A. Zhang, H. Ji, and T. Abdelzaher. (2021). "Controllable and diverse text generation in e-commerce". In: *Proceedings of the Web Conference 2021*. 2392–2401.

Shen, S., C. Li, X. Hu, Y. Xie, J. Yang, P. Zhang, Z. Gan, L. Wang, L. Yuan, C. Liu, *et al.* (2022). "K-lite: Learning transferable visual models with external knowledge". *Advances in Neural Information Processing Systems*. 35: 15558–15573.

Shi, Y., B. Paige, P. Torr, *et al.* (2019). "Variational mixture-of-experts autoencoders for multi-modal deep generative models". *Advances in neural information processing systems*. 32.

Shilova, V., L. D. Santos, F. Vasile, G. Racic, and U. Tanielian. (2023). "AdBooster: Personalized Ad Creative Generation using Stable Diffusion Outpainting". *arXiv preprint arXiv:2309.11507*.

Sohl-Dickstein, J., E. Weiss, N. Maheswaranathan, and S. Ganguli. (2015). "Deep unsupervised learning using nonequilibrium thermodynamics". In: *International conference on machine learning*. PMLR. 2256–2265.

Song, J., C. Meng, and S. Ermon. (2020a). "Denoising diffusion implicit models". *arXiv preprint arXiv:2010.02502*.

Song, Y., J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. (2020b). "Score-based generative modeling through stochastic differential equations". *arXiv preprint arXiv:2011.13456*.

Su, J. (2018). "GAN-QP: A Novel GAN Framework without Gradient Vanishing and Lipschitz Constraint". *ArXiv*. abs/1811.07296. URL: https://api.semanticscholar.org/CorpusID:53713813.

Sun, Y., H. Xue, R. Song, B. Liu, H. Yang, and J. Fu. (2022). "Long-form video-language pre-training with multimodal temporal contrastive learning". *Advances in neural information processing systems*. 35: 38032–38045.

Sutter, T., I. Daunhawer, and J. Vogt. (2020). "Multimodal generative learning utilizing jensen-shannon-divergence". *Advances in neural information processing systems*. 33: 6100–6110.

Tang, Z., Z. Yang, M. Khademi, Y. Liu, C. Zhu, and M. Bansal. (2023). "CoDi-2: In-Context, Interleaved, and Interactive Any-to-Any Generation". arXiv: 2311.18775 [cs.CV].

Tang, Z., Z. Yang, C. Zhu, M. Zeng, and M. Bansal. (2024). "Any-to-any generation via composable diffusion". *Advances in Neural Information Processing Systems*. 36.

Tautkute, I. and T. Trzcinski. (2021). "I want this product but different: Multimodal retrieval with synthetic query expansion". *arXiv preprint arXiv:2102.08871*.

Team, A. (2024). "Introducing the next generation of Claude". https://www.anthropic.com/news/claude-3-family.

Team, G., R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, *et al.* (2023). "Gemini: a family of highly capable multimodal models". *arXiv preprint arXiv:2312.11805*.

Templeton, S. (2024). "What Is Shop with Google AI and Why Should I Care?" https://www.singlegrain.com/blog/n/shop-with-google-ai/.

Touvron, H., L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, *et al.* (2023). "Llama 2: Open foundation and fine-tuned chat models". *arXiv preprint arXiv:2307.09288*.

Truong, Q.-T. and H. Lauw. (2019). "Multimodal review generation for recommender systems". In: *The World Wide Web Conference*. 1864–1874.

Vyas, A., B. Shi, M. Le, A. Tjandra, Y.-C. Wu, B. Guo, J. Zhang, X. Zhang, R. Adkins, W. Ngan, *et al.* (2023). "Audiobox: Unified audio generation with natural language prompts". *arXiv preprint arXiv:2312.15821*.

Wang, H., W. Feng, Y. Lu, Y. Li, Z. Zhang, J. Lv, X. Zhu, J. Shen, Z. Lin, L. Bo, *et al.* (2023). "Generate E-commerce Product Background by Integrating Category Commonality and Personalized Style". *arXiv preprint arXiv:2312.13309*.

Wang, L., P. Luc, Y. Wu, A. Recasens, L. Smaira, A. Brock, A. Jaegle, J.-B. Alayrac, S. Dieleman, J. Carreira, *et al.* (2022a). "Towards learning universal audio representations". In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 4593–4597.

Wang, R., H. Guo, J. Liu, H. Li, H. Zhao, X. Tang, Y. Hu, H. Tang, and P. Li. (2024a). "StableGarment: Garment-Centric Generation via Stable Diffusion". *arXiv preprint arXiv:2403.10783*.

Wang, T. and M. Ye. (2024). "TexFit: Text-Driven Fashion Image Editing with Diffusion Models". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. No. 9. 10198–10206.

Wang, X., B. Zhuang, and Q. Wu. (2024b). "ModaVerse: Efficiently Transforming Modalities with LLMs". *arXiv preprint arXiv:2401.06395*.

Wang, Y., Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi. (2022b). "Self-instruct: Aligning language models with self-generated instructions". *arXiv preprint arXiv:2212.10560*.

Wang, Y., H. Zhang, Z. Liu, L. Yang, and P. S. Yu. (2022c). "Contrastvae: Contrastive variational autoencoder for sequential recommendation". In: *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 2056–2066.

Watson, D., W. Chan, J. Ho, and M. Norouzi. (2021). "Learning fast samplers for diffusion models by differentiating through sample quality". In: *International Conference on Learning Representations*.

Wei, P., S. Liu, X. Yang, L. Wang, and B. Zheng. (2022). "Towards personalized bundle creative generation with contrastive non-autoregressive decoding". In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2634–2638.

Wei, W., C. Huang, L. Xia, and C. Zhang. (2023). "Multi-Modal Self-Supervised Learning for Recommendation". In: *Proceedings of the ACM Web Conference 2023*. 790–800.

Wei, Y., X. Wang, L. Nie, X. He, R. Hong, and T.-S. Chua. (2019). "MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video". In: *Proceedings of the 27th ACM international conference on multimedia*. 1437–1445.

Won, M., S. Chun, O. Nieto, and X. Serrc. (2020). "Data-driven harmonic filters for audio representation learning". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 536–540.

Wu, L., Z. Zheng, Z. Qiu, H. Wang, H. Gu, T. Shen, C. Qin, C. Zhu, H. Zhu, Q. Liu, *et al.* (2023a). "A survey on large language models for recommendation". *arXiv preprint arXiv:2305.19860*.

Wu, M. and N. Goodman. (2018). "Multimodal generative models for scalable weakly-supervised learning". *Advances in neural information processing systems*. 31.

Wu, S., H. Fei, L. Qu, W. Ji, and T.-S. Chua. (2023b). "Next-gpt: Any-to-any multimodal llm". *arXiv preprint arXiv:2309.05519*.

Xiao, F., L. Deng, J. Chen, H. Ji, X. Yang, Z. Ding, and B. Long. (2022). "From abstract to details: A generative multimodal fusion framework for recommendation". In: *Proceedings of the 30th ACM International Conference on Multimedia*. 258–267.

Xie, J., K. Zhang, J. Chen, T. Zhu, R. Lou, Y. Tian, Y. Xiao, and Y. Su. (2024). "Travelplanner: A benchmark for real-world planning with language agents". *arXiv preprint arXiv:2402.01622*.

Xu, Y., W. Wang, F. Feng, Y. Ma, J. Zhang, and X. He. (2024a). "DiFashion: Towards Personalized Outfit Generation". *arXiv preprint arXiv:2402.17279*.

Xu, Y., T. Gu, W. Chen, and C. Chen. (2024b). "OOTDiffusion: Outfitting Fusion based Latent Diffusion for Controllable Virtual Try-on". *arXiv preprint arXiv:2403.01779*.

Yan, A., Z. He, J. Li, T. Zhang, and J. McAuley. (2023). "Personalized showcases: Generating multi-modal explanations for recommendations". In: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2251–2255.

Yang, D., J. Yu, H. Wang, W. Wang, C. Weng, Y. Zou, and D. Yu. (2023a). "Diffsound: Discrete diffusion model for text-to-sound generation". *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Yang, J., J. Duan, S. Tran, Y. Xu, S. Chanda, L. Chen, B. Zeng, T. Chilimbi, and J. Huang. (2022). "Vision-language pre-training with triple contrastive learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15671–15680.

Yang, L., Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang. (2023b). "Diffusion models: A comprehensive survey of methods and applications". *ACM Computing Surveys*. 56(4): 1–39.

Yang, R., L. Song, Y. Li, S. Zhao, Y. Ge, X. Li, and Y. Shan. (2024). "Gpt4tools: Teaching large language model to use tools via self-instruction". *Advances in Neural Information Processing Systems*. 36.

Yi, J., Y. Zhu, J. Xie, and Z. Chen. (2021). "Cross-modal variational auto-encoder for content-based micro-video background music recommendation". *IEEE Transactions on Multimedia*. 25: 515–528.

Yi, Z., X. Wang, I. Ounis, and C. Macdonald. (2022). "Multi-modal graph contrastive learning for micro-video recommendation". In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1807–1811.

Yin, S., C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen. (2023). "A survey on multimodal large language models". *arXiv preprint arXiv:2306.13549*.

Yu, H., Y. Tian, S. Kumar, L. Yang, and H. Wang. (2023). "The devil is in the details: A deep dive into the rabbit hole of data filtering". *arXiv preprint arXiv:2309.15954*.

Yuan, M., I. R. Khan, F. Farbiz, S. Yao, A. Niswar, and M.-H. Foo. (2013). "A mixed reality virtual clothes try-on system". *IEEE Transactions on Multimedia*. 15(8): 1958–1968.

Zhai, X., X. Wang, B. Mustafa, A. Steiner, D. Keysers, A. Kolesnikov, and L. Beyer. (2022). "Lit: Zero-shot transfer with locked-image text tuning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18123–18133.

Zhang, D., Y. Yu, C. Li, J. Dong, D. Su, C. Chu, and D. Yu. (2024). "Mm-llms: Recent advances in multimodal large language models". *arXiv preprint arXiv:2401.13601*.

Zhang, H., Y. Zou, and H. Wang. (2021). "Contrastive self-supervised learning for text-independent speaker verification". In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 6713–6717.

Zhang, L., A. Rao, and M. Agrawala. (2023). "Adding conditional control to text-to-image diffusion models". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3836–3847.

Zhang, Z., C. Luo, and J. Yu. (2019). "Towards the Gradient Vanishing, Divergence Mismatching and Mode Collapse of Generative Adversarial Nets". In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management. CIKM '19*. Beijing, China: Association for Computing Machinery. 2377–2380. ISBN: 9781450369763. DOI: 10.1145/3357384.3358081. URL: https://doi.org/10.1145/3357384.3358081.

Zhou, D., H. Zhang, J. Ma, and J. Shi. (2023a). "BC-GAN: A Generative Adversarial Network for Synthesizing a Batch of Collocated Clothing". *IEEE Transactions on Circuits and Systems for Video Technology*.

Zhou, X. and C. Miao. (2024). "Disentangled Graph Variational Auto-Encoder for Multimodal Recommendation With Interpretability". *IEEE Transactions on Multimedia*.

Zhou, Z., Y. Lei, B. Zhang, L. Liu, and Y. Liu. (2023b). "Zegclip: Towards adapting clip for zero-shot semantic segmentation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11175–11185.

Zhu, L., D. Yang, T. Zhu, F. Reda, W. Chan, C. Saharia, M. Norouzi, and I. Kemelmacher-Shlizerman. (2023). "Tryondiffusion: A tale of two unets". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4606–4615.

Zhu, Z., Y. Li, W. Lyu, K. K. Singh, Z. Shu, S. Pirk, and D. Hoiem. (2024). "Consistent Multimodal Generation via A Unified GAN Framework". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 5048–5057.

Ziegler, J. D., S. Subramaniam, M. Azzarito, O. Doyle, P. Krusche, and T. Coroller. (2022). "Multi-modal conditional GAN: Data synthesis in the medical domain". In: *NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research*.