C3Net: Compound Conditioned ControlNet for Multimodal Content Generation

Juntao Zhang[1*] Yuehuai Liu[1*] Yu-Wing Tai[2] Chi-Keung Tang[1]
[1]HKUST [2]Dartmouth College

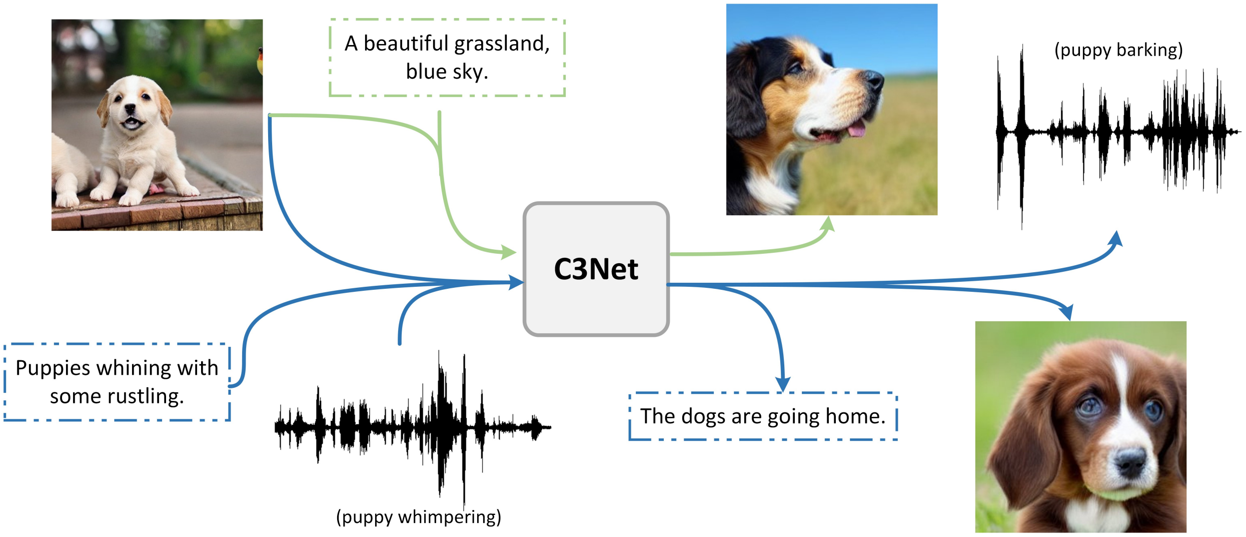
A beautiful grassland, blue sky.
(puppy barking)
C3Net
Puppies whining with some rustling.
The dogs are going home.
(puppy whimpering)

Figure 1. **C3Net** generates multimodal contents (e.g., image, audio, and text) taking compound conditions in multiple modalities. The green and blue arrows are two inferences made by C3Net using different combinations of conditions. C3Net can take any combination of image, text, and audio as conditions for content synthesis.

Abstract

We present Compound Conditioned ControlNet, C3Net, a novel generative neural architecture taking conditions from multiple modalities and synthesizing multimodal contents simultaneously (e.g., image, text, audio). C3Net adapts the ControlNet [46] architecture to jointly train and make inferences on a production-ready diffusion model and its trainable copies. Specifically, C3Net first aligns the conditions from multi-modalities to the same semantic latent space using modality-specific encoders based on contrastive training. Then, it generates multimodal outputs based on the aligned latent space, whose semantic information is combined using a ControlNet-like architecture called Control C3-UNet. Correspondingly, with this system design, our model offers an improved solution for joint-modality generation through learning and explaining multimodal conditions, involving more than just linear interpolation within the latent space. Meanwhile, as we align conditions to a unified latent space, C3Net only requires one trainable Control C3-UNet to work on multimodal semantic information. Furthermore, our model employs unimodal pretraining on the condition alignment stage, outperforming the non-pretrained alignment even on relatively scarce training data and thus demonstrating high-quality compound condition generation. We contribute the first high-quality tri-modal validation set to validate quantitatively that C3Net outperforms or is on par with the first and contemporary state-of-the-art multimodal generation [43]. Our codes and tri-modal dataset will be released here.

1. Introduction

Diffusion models have recently emerged as the new state-of-the-art family of deep generative models, with remarkable performance on multimodal modeling [1, 33, 35, 37, 47]. Correspondingly, we have observed widespread and increasing prevalence of strong cross-modal models that allow generating one single modality from another, including but not limited to text-to-text [2, 30], text-to-

[*] Equal contributions. This work is supported in part by the Research Grant Council of the Hong Kong SAR grant no. 16201420.

image [5, 12, 13, 35, 39] and text-to-audio [14, 25]. However, these existing models cannot simultaneously accept a wider range of input modalities than text or image, nor are they capable of simultaneously generating multiple output modalities in parallel, which leads to limited application in most real-world scenarios where multiple modalities coexist and overlap with one another. The generation capability of each step remains intrinsically constrained even when modality-specific generative models are chained in sequence a multi-step generation setup, which can be laborious, time-consuming and compute-demanding. In this regard, Composable Diffusion (CoDi) [43] is to date the only contemporary work capable of concurrently generating any combinations of modalities, simply by taking linear interpolations on the aligned latent space, which results in the downgraded synthesis qualities. Thus, a better and more flexible joint-modality generative model is necessary.

To achieve better synthesis results while facilitating "any-to-any" generation capabilities, we propose Compound Conditioned ControlNet, or *C3Net*, whose overall architecture design is adapted from ControlNet [46], which trains and makes inferences on a production-ready diffusion model and its trainable copies. Our model first aligns the conditions obtained from individual modalities to a shared semantic latent space. During the training of alignment encoders, we utilize unimodal pre-training to mitigate the deficiency of high-quality multimodal datasets. The semantic information obtained from individual modalities is further combined through a learnable ControlNet-like architecture called Control C3-UNet. multimodal conditions are then coordinated and merged into the C3-UNet for multimodal synthesis. Consequently, our model can generate multimodal outputs from the aligned latent space.

Thus, *1)* C3Net contributes a better solution than straightforward linear interpolations on the latent space, synthesizing more complex and diverse outputs beyond them. Notably, C3Net only requires training one Control C3-UNet to work on multimodal conditions, which substantially reduces complexity for joint-modality training and generation. Furthermore, *2)* C3Net employs unimodal pre-training on the condition alignment stage, which facilitates alignment quality even on relatively scarce training data. Overall, C3Net outperforms or is on par with the state-of-the-art multimodal generation counterparts, making it the next strong baseline for generating complex and diverse multimodal outputs.

2. Related Work

2.1. Diffusion Models

Diffusion models (DMs) consist of a class of probabilistic generative models capable of understanding the desired data distribution and synthesizing new samples, through a continuous application of denoising autoencoders in output generation. For the three dominant formulations, *Denoising diffusion probabilistic models (DDPMs)* [11] utilize two Markov Chains for image generation: a forward chain that injects random noise to the data and transforms the data distribution into an unstructured simple prior, and a reverse chain that denoises and recovers the original data by understanding the learnable transition kernels. *Score-based generative models (SGMs)* [40, 41] introduce score functions defined as the gradient of log probability density, adding a series of escalating Gaussian noise into the data and jointly calculating the scores for all noisy data distributions. *Stochastic Differential Equations (SDEs)* [41] can further be leveraged in the injection and denoising processes, allowing for the scenario of unlimited time steps or noise levels in DDPMs and SGMs. *Latent diffusion models (LDMs)* [35] first train a *Variational autoencoder (VAE)* [18, 34] to encode inputs into a low-dimensional and efficient latent space, and then apply a diffusion model to further generate latent codes. By abstracting negligible details and reducing modeling dimension, the motivation is to focus on the semantic aspects of the data to achieve higher computational efficiency. The diffusion models have achieved state-of-the-art synthesis quality in image inpainting, image superresolution, and audio generation from text.

2.2. Composable Diffusion

Composable Diffusion (CoDi) [43] is a joint-modality generative model capable of producing a combination of output modalities in parallel based on a combination of input modalities, such as text, audio, image and video. CoDi first trains a latent diffusion model for each modality independently, adds a cross-attention module to each diffuser, and further apply an environment encoder to project the latent variables of different LDMs into a shared latent space. CoDi's design enables multimodal generation without training on all possible combinations of modalities, reducing the size of training from one of exponential to linear.

2.3. Unimodal Pre-training

Unimodal Models trained on large single-modality datasets can achieve a broader and more diverse coverage of real-world data distribution, without being constrained by the presence of cross-modality data pairs. Specifically, using unimodal models as pre-training can achieve better zero-shot performance compared with the jointly-trained multimodal models, with *MAE* [8] and *T5* [32] outperforming the state-of-the-art CLIP-based model under similar model capacities. Moreover, as an effective unimodal pre-training technique for audio processing tasks, *Self-Supervised Audio Spectrogram Transformer (SSAST)* [7] enables models to learn the underlying patterns and features of large, unlabeled audio datasets and further improve their performance

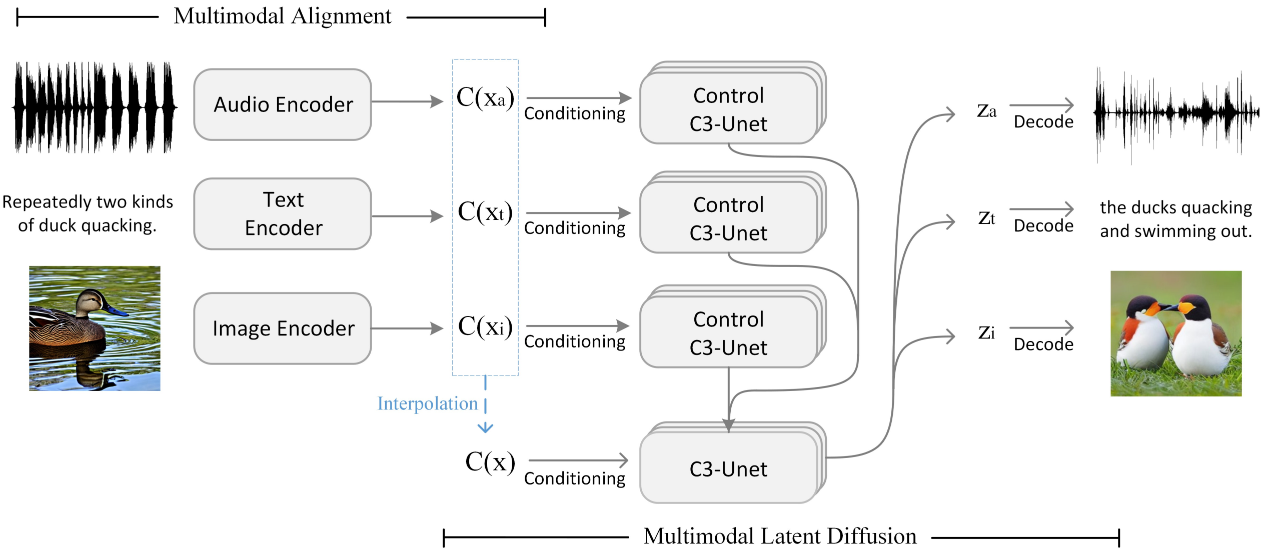

Figure 2. **C3Net** first aligns compound conditions in different modalities to a shared latent space ξ, where the encoder takes individuals of compound conditions and generates aligned latent: $C(x_a)$, $C(x_t)$, $C(x_i)$ are all in ξ, and $C(x)$ is an interpolation. The aligned condition latent is fed to a generative network consisting of *C3-UNet* and *Control C3-UNet*, which adaptively learns to coordinate compound conditions in addition to the weighted arithmetic mean in the latent space indicated by the blue dashed box. Thanks to the common latent ξ, the Control C3-Unet for different conditions shares the same weight. C3Net generates multiple latent for each modality, denoted as z_a, z_t, z_i, for audio, text, and image, respectively. Then, the z's are decoded using their respective established decoders to generate contents. See Figure 4 for Control C3-UNet and C3-UNet details.

on the fine-tuning datasets. In the case of C3Net, we first apply unimodal pre-trained encoders for each modality, and then fine-tune the encoders on smaller-scale high-quality datasets based on contrastive learning.

2.4. Multimodal Alignment

Contrastive Language-Image Pre-Training (CLIP) [31, 33] is a neural network that aligns the text and image modalities by pre-training on a large dataset of text-image pairs with a contrastive loss function. Given a sample size of N text-image pairs, CLIP learns to map the two modalities into a common embedding space by jointly training a text encoder and image encoder to maximize the cosine similarity of N matched pairs and minimize the cosine similarity of $(N^2 - N)$ unmatched pairs using a contrastive loss function. Similar with CLIP, *Contrastive Language-Audio Pre-Training (CLAP)* [4] aligns the text and audio modalities via contrastive learning paradigm between the audio and text embeddings in pair, also following the same loss function. *CoDi* [43] proposes the “Bridging Alignment” technique to align conditional encoders for multimodal generation. CoDi leverages CLIP as the pretrained text-image paired encoder, and trains audio and video prompt encoders on audio-text and video-text paired datasets using contrastive learning, with text and image encoder weights frozen.

The above alignment techniques can be applied to align the latent space of LDMs with multiple modalities to achieve joint multimodal generation. In comparison, C3Net also utilizes the modality-specific encoders to align the conditions from multi-modalities to the same latent space, while it takes a step further by adding neural architecture similar with ControlNet [46] to facilitate better understanding of multimodal conditions and joint-modality generation.

2.5. ControlNet

ControlNet [46] learns and adds spatial conditioning to control the large pre-trained diffusion models. By freezing the original weights for the pre-trained diffusion model, ControlNet leverages a trainable copy of its deep-and-robust encoding layers to learn the diverse set of conditional controls and avoid overfitting. The original locked model and the trainable copy are then connected with a zero-initialized convolution layer called “zero-convolution,” where the convolution weights are first initialized to zero and progressively learned throughout the training.

This architecture provides an effective solution for controlling large diffusion models, while ensuring that no new and harmful noises would be added to the deep features of the diffusion models. In the design of C3Net, we independently apply a ControlNet-like architecture to each input modality, further enabling our model to learn to coordinate multimodal conditions and synthesize more optimized results in the cross-modality generation.

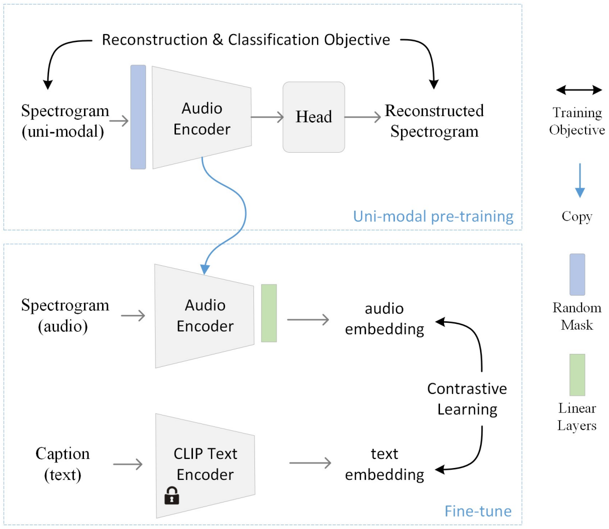

Figure 3. **Uni-modal pre-training** of audio encoder. The audio encoder is first initialized using unsupervised pre-training on large-scale uni-modal data. The encoder is then fine-tuned with objective learning using high-quality multi-modal data.

3. Method

C3Net is a neural network architecture for synthesizing multimodal content conditioned on multimodal inputs. Figure 2 shows C3Net's overall architecture with content in different modalities (audio, text and image). We first introduce C3Net's overall structure in Section 3.1, including the alignment encoder $C(\cdot)$, and the latent diffusion model consisting of *Control C3-Unet* and *C3-Unet*. Then, we explain the unimodal pre-training for encoders $C(\cdot)$ in Section 3.2 where, unlike many existing multimodal approaches, multimodal training data (e.g., audio and image pair-up) is not needed in the pre-training stage. Finally, we explain *C3-UNet* and *Control C3-UNet* for compound conditional generation in Section 3.3.

3.1. Compound Conditioned ControlNet

C3Net (Compound Conditioned ControlNet) takes inspiration of the general architecture from [43] to enable multimodal generation conditioned on compound information, such as audio, text and image. C3Net first aligns multimodal conditions to a shared latent space. We consider a compound multimodal condition c_a, c_i, c_t, respectively denoting audio, image, and text conditions, and project them to a shared latent space ξ using an encoder $C(\cdot)$.

We observed that text captioning is ubiquitously adapted in large-scale multimodal datasets as one of the ground truth labels, and that as mentioned in [43], certain dual modalities datasets (e.g., audio and image) are either harshly-aligned or scarce in quantity. To address this issue, choosing a shared latent space ξ in which the text encoder is well-established is advisable. Following the practical implementation of [43], we adopt CLIP [31] latent space as our shared latent space ξ. Thus, an instance of the compound condition alignment stage yields a tuple of latent

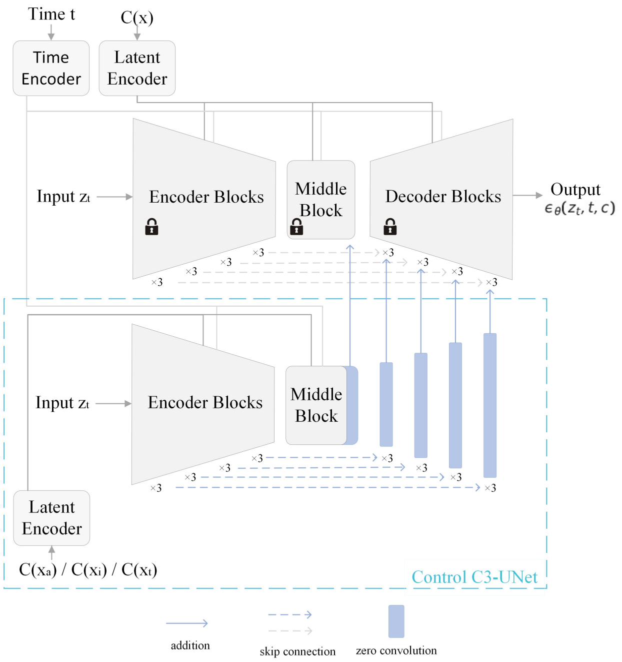

Figure 4. Multimodal generation of C3Net consists of **C3-UNet** (top) and **Control C3-UNet** (bottom in the blue box). Similar to ControlNet, Control C3-UNet provides additional control to the C3-UNet. Control C3-UNet takes latent condition aligned from each modality separately and connects to the C3-UNet at each level of skip-connection by addition. $C(x_a)$, $C(x_i)$, $C(x_t)$ are aligned audio, image, text conditions respectively; they are input one at a time. Each modality is generated by its respective UNet pair.

$$C(x_a), C(x_i), C(x_t) \in \xi \tag{1}$$

denoting the aligned latent from audio, image, and text conditions, respectively.

After acquiring latent conditions, we generate multimodal contents using latent diffusion models with C3-UNet and Control C3-UNet as the backbone. Specifically, we can sample feature maps z_0 of any modality from a diffusion model sampling process, which is conditioned on $C(x_a), C(x_i), C(x_t)$. Note that the synthesis of different modalities utilizes different diffusion models. Then, z_0 is decoded on respective decoders to generate the content of its modality.

3.2. Uni-modal Pre-trained Alignment

Our encoders $C(\cdot)$ are multi-modal encoders aligning conditions in different modalities to the shared space ξ. The encoders are first pre-trained on unimodal datasets and then fine-tuned using contrastive learning proposed in [31]. In contrast, the original settings of [43] is an alignment model trained from scratch on multi-modal datasets.

We propose to pre-train encoders on *unimodal* data because high-quality paired datasets are scarce for some modalities (e.g., audio and text pair). On the other hand, many high-quality unimodal datasets are readily available. In the following, we use the audio encoder as an example. As shown in Figure 3, we use the pre-trained neural net from [7] which is a masked auto-encoder. The MAE has been trained to extract audio features during the unsupervised training stage, which makes it easier for the following contrastive learning for the audio encoder. We then fine-tune the audio encoder using high-quality datasets, which are available on relatively small scales. During the fine-tuning stage of the audio encoder, we utilize an established frozen text encoder from [31]. In detail, we use contrastive objective to fine-tune the pre-trained audio encoder, so that the audio encoder learns to align audio to latent in ξ as similar as possible to the latent that its ground truth caption is aligned to.

Similar to the findings in [20–22, 27, 42, 48], our encoder networks can primarily learn the data pattern for respective modalities with unsupervised pre-training. Trained with fewer but high-quality multi-modal data, our unimodal pre-trained encoder is on par with or outperformed encoders trained on only paired data on downstream generation tasks.

3.3. C3-UNet and Control C3-UNet

Figure 4 shows the multimodal diffusion model of our C3Net, which consists of the C3-UNet and Control C3-UNet. *C3-UNet* is a trained UNet $\mathcal{F}(\cdot, \Theta)$ employed in a latent diffusion model, which generates feature maps conditioned on instances in ξ. *Control C3-UNet*, similar to the ControlNet setting in [46], is a trainable copy $\mathcal{F}(\cdot, \Theta_c)$ of the C3-UNet, where Θ_c denotes the trainable copy of parameters Θ. In the implementation of C3Net, we use the trained UNet of Composable Diffusion [43] as C3-UNet $\mathcal{F}(\cdot, \Theta)$. Figure 4 shows the detailed architecture[1].

The Control C3-UNet can provide additional information lost during the latent interpolation. Notably, our Control C3Net takes the aligned latent $C(x_a), C(x_i)$, and $C(x_t)$ separately in each modality. The Control C3-UNet is trained to provide extra information in each condition by modifying the feature maps of the C3-UNet. Thanks to the shared latent space ξ, it is sufficient to train *one trainable copy of parameters* Θ_c for the Control C3-UNet. This is because conditions from all modalities have been aligned to the shared ξ, and a single set of trained parameters Θ_c is sufficient for additional control by taking condition latent already aligned in ξ regardless of the original modality.

Our diffusion model follows a similar setting in [43] and [31]. The C3-UNet takes a linear interpolation of the aligned latent $C(x_a), C(x_i), C(x_t)$ as the condition. The Control C3-UNet takes conditions in each modality separately and connects to the C3-UNet at each level of skip-connection after multiplying a constant, which we empirically set to be 0.1, but it varies depending on the generation task. Constant multiplied adjusts the additional control scale the Control C3-UNet provides when using different combinations of compound conditions. However, when only a single condition is provided, the Control C3-UNet can not provide additional information and therefore we set the constant to zero.

During training, we use the text-image dataset to train C3Net's image and text generation, and the text-audio dataset for audio generation (training and validation datasets will be described in Section 4.1). Specifically, we train Control C3-UNet on each modality separately. Take image generation as an example, for each image-text pair, denoted as I and x respectively, in the dataset. The ground truth text x is aligned to the shared latent space as $C(x)$, and a masked text x_m is generated by randomly selecting 50% of the text prompt to replace with empty strings and it is aligned as $C(x_m)$. The C3-UNet takes $C(x_m)$, and Control C3-UNet takes $C(x)$ as condition latent, respectively. The ground truth image I is used to generate z_0 in a typical latent diffusion model [35]. We train the C3-UNet to predict noise in a timestep t during the image diffusion. Therefore, the objective function of each modality can be denoted as

$$\mathcal{L}_c = \mathbb{E}_{z_0,t,c,\epsilon \in \mathcal{N}(0,1)} \left[\|\epsilon - \epsilon_\theta(z_t, t, c)\|_2^2 \right] \tag{2}$$

where ϵ is the ground truth noise, $\epsilon_\theta(\cdot)$ is the network, and c is the tuple of $C(x_m), C(x)$.

4. Experiments

4.1. Training Datasets

We collected our training datasets for fine-tuning the alignment encoder as well as the respective Control C3-UNet for image, audio, and text synthesis. Major effort was made to clean up flawed data in some datasets through data prepossessing and relation scores, including CLIP score [9], CLAP [4] similarity score, and the data quality metrics.

[1] The trained $\mathcal{F}(\cdot, \Theta)$ is a U-Net with an encoder, a middle block, and a skip-connected decoder. The encoder and decoder contain 12 blocks, and the full model contains 25 blocks, including the middle block. Of the 25 blocks, there are 4 down-sampling and 4 up-sampling blocks. Refer to Figure 4. In C3-UNet, the "Encoder Blocks" contains 12 encoder blocks in 4 resolutions, while the "×3" indicates the block of the same resolution repeats three times. Condition latent is encoded using the Latent Encoder, and diffusion time steps are encoded with a time encoder using positional encoding. Similar to [46], the Control C3-UNet is a trainable copy of the 12 encoder block and 1 middle block of the C3-UNet. The feature maps are added to the 12 skip-connections and 1 middle block of the C3-UNet after a "zero convolution" layer.

For the fine-tuning of *audio encoder*, we used AudioCap [17], a dataset of sounds with event descriptions for audio captioning. We also added the sound effects from Epidemic Sound as provided in [45]. We selected 1 million ten-second sound clips from AudioSet [6] with optimal quality and CLAP similarity score to their text captions.

For the fine-tuning of *image and text Control C3-UNets*, we utilized the COCO [24] dataset and part of the LAION-400M [38] dataset, with both consisting of images and corresponding text captions. For the fine-tuning of *audio Control C3-UNet*, we utilized a combination of AudioCap [17] and AudioSet [6] for training.

4.2. Tri-modality Test Set

In the absence of k-modality datasets, where $k > 2$, for multimodal synthesis evaluation, it is crucial to construct a high-quality evaluation set with three modalities (i.e., image, text, and audio) for evaluating C3Net. It is important to note that a bi-modal test set would be unsuitable for C3Net, as it only provides one modality as a condition (with the other serving as ground truth) and cannot effectively evaluate the performance of multimodal conditioned synthesis.

Observing that the AudioCap dataset [17] contains high-quality audio and text captions, we curated a tri-modal test set based on AudioCap. Specifically, we first generated the third modality (i.e., image) using Stable Diffusion [36] prompted on the AudioCap text captions. We further selected 2,000 data tuples based on image quality and CLIP score to ensure that the content for each tuple is highly correlated. As a result, a total of 2,000 tri-modal ground-truth tuples, each with highly relevant audio, image and text caption, are available for evaluating C3Net and CoDi [43], which is to date the most representative (and only) work on diffusion-based tri-modality content generation. We show some examples in Figure 5.

4.3. Evaluation Results

Figure 6 shows qualitative comparisons on compound condition image synthesis between C3Net and our baseline. We will show quantitative comparison in the following.

4.3.1 Unimodal Pre-training

We evaluate the unimodal pre-training results by comparing the image and text generated by C3Net and CoDi [43]. Evaluation is conducted on the AudioCap [17] test set, with which we generate text captions and images conditioned on the ground truth audio. Table 1 shows the correlation between the generated text and its ground truth captions, assessed by scaled CIDEr-D [44] and SPIDEr [26]. Table 2 tabulates the image synthesis quality assessed by the Inception Score, as well as the correlation between the generated images and ground truth text captions assessed by the CLIP score [9]. Evaluations on audio synthesis are not available in this case, as the model design of CoDi does not support audio generation when taking audio as a condition. Note that in the scenario of taking an audio as the only condition, C3Net and CoDi differ only in the alignment stage, which makes it an ideal ablation study. Under such settings, C3Net applies an audio encoder pre-trained on unimodal data, while CoDi uses an audio encoder without unimodal pre-training.

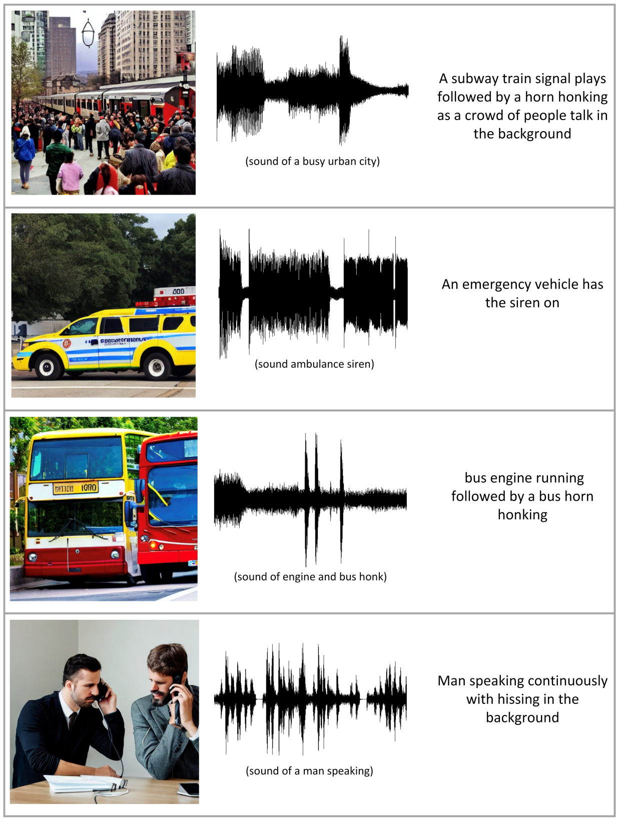
(sound of a busy urban city)
A subway train signal plays followed by a horn honking as a crowd of people talk in the background
(sound ambulance siren)
An emergency vehicle has the siren on
(sound of engine and bus honk)
bus engine running followed by a bus horn honking
(sound of a man speaking)
Man speaking continuously with hissing in the background

Figure 5. Examples from the Tri-modality Test Set.

Method	CIDEr-D ↑	SPIDEr ↑
CoDi	0.0654	0.0608
C3Net (Ours)	**0.0704**	**0.0622**

Table 1. Unimodal pre-training assessed by the correlation between the synthesized **texts** and ground truth text captions on the AudioCap test set. Comparisons are made between C3Net (with unimodal pre-trained encoders) and CoDi (without as such).

4.3.2 Multimodal Synthesis

We evaluate the multimodal synthesis capabilities of C3Net on the Tri-modality Test Set introduced in 4.2. To assess the synthesis quality on compound conditions, we generate

Method	Inception Score ↑	CLIP ↑
CoDi	1.7730, 0.1450	23.192
C3Net (Ours)	**1.7732**, **0.1535**	**23.325**

Table 2. Unimodal pre-training assessed by Inception Score of the generated **images**, and the CLIP score between the generated **images** and the ground truth text captions. Images are synthesized conditioning on the AudioSet test set audio.

images, text, and audio conditioned on each respective tuple within the Tri-modality Test Set.

To evaluate **image** synthesis, we measure the Fréchet inception distance [10] between the synthesized image and its ground truth image as well as the CLIP score [9] between the generated image and its ground truth text caption. Table 3 shows that C3Net generates images that relate closer to both the text and image conditions, demonstrating that our Control C3-UNet architecture offers a more optimized solution for compound condition image synthesis.

Method	FID ↓	CLIP ↑
CoDi	11.39	25.17
C3Net (Ours)	**10.97**	**25.29**

Table 3. Compound conditioned **image** synthesis assessed on the Tri-modality Test Set. Generation quality is measured by the Fréchet inception distance between the synthesized image and its ground truth image, and the CLIP score between the synthesized image and its ground truth text caption.

To evaluate **text** synthesis, we measure the caption correlation metrics between the synthesized text and the ground truth captions. The caption metrics include BLEU-1 [28], ROUGE-L [23], CIDEr-D [44], and SPIDEr [26]. Table 4 shows that C3Net generates text outputs more closely correlated with the ground truth compared to CoDi.

Method	BLEU-1 ↑	ROUGE-L ↑	CIDEr-D ↑	SPIDEr ↑
CoDi	0.1059	0.1019	0.0651	0.0631
C3Net (Ours)	**0.1104**	**0.1045**	**0.0713**	**0.0665**

Table 4. **Text** synthesis assessed on the Tri-modality Test Set. We evaluate the correlation between synthesized texts and ground truth text captions using a variety of caption metrics.

To evaluate **audio** synthesis, we measure OVL (Overall Impression), REL (Text Relevant) similar with the settings in [19], and FAD [16] to evaluate the audio quality. As shown in Table 5, C3Net outperforms CoDi in terms of OVL and REL. When measuring the correlation between the synthesized audio and the ground truth audio, C3Net yields a slightly weaker FAD score compared to CoDi.

Method	OVL ↑	REL ↑	FAD ↓
Reference	81.07	79.31	-
CoDi	62.91	59.01	11.4
C3Net (Ours)	**63.25**	**59.83**	11.7

Table 5. **Audio** synthesis assessed on the Tri-modality Test Set. The audio evaluation metrics include OVL and REL between the synthesized audios and the ground truth captions. We also evaluated the FAD between generated audio and its ground truth.

4.3.3 Synthesized Audio Classification

To further assess the quality of audio synthesized by C3Net, we compare the classification accuracy of the generated image conditioned on the audio-text pairs in the ESC-50 [29] dataset. In this experiment, we first synthesized audio conditioned on the audio and text pairs. Then, we classified the generated audio using the classification model given in [3]. Table 6 tabulates the results, where a higher accuracy indicates more optimized audio synthesis on compound conditions, which keeps the shared features in multimodal conditions.

	Codi	C3Net (Ours)
Accuracy (%)	21.05	**23.25**

Table 6. Classification accuracy on synthesized **audio** conditioned on audio-text pairs in ESC-50. A higher accuracy indicates a better ability to keep shared features in multimodal conditions.

5. Conclusion and Discussion

In this paper, we propose C3Net, a multimodal generative model conditioned on compound content, which applies unsupervised pre-training on unimodal datasets and further leverages a ControlNet-like architecture to coordinate compound conditions. Through extensive experiments, we demonstrate that C3Net is capable of synthesizing high-quality multimodal contents on compound conditions by coordinating them through a learnable process, and addressing the deficiencies of datasets through unimodal pre-training.

While C3Net has shown remarkable progress in joint-modality generation, there exist remaining challenges that need to be addressed in the future. One of the issues is the choice of the shared latent space, such as the CLIP [31] latent, which may not be optimal for all modalities, particularly audio. To address this issue, a contrastive learning process that takes into account multiple modalities may be more effective. Another challenge is that aligning latent conditions using contrastive learning may sacrifice the unique information contained in a modality, as noted in a previous study [15]. One solution to this issue is to use a similar alignment objective as proposed in [15], which

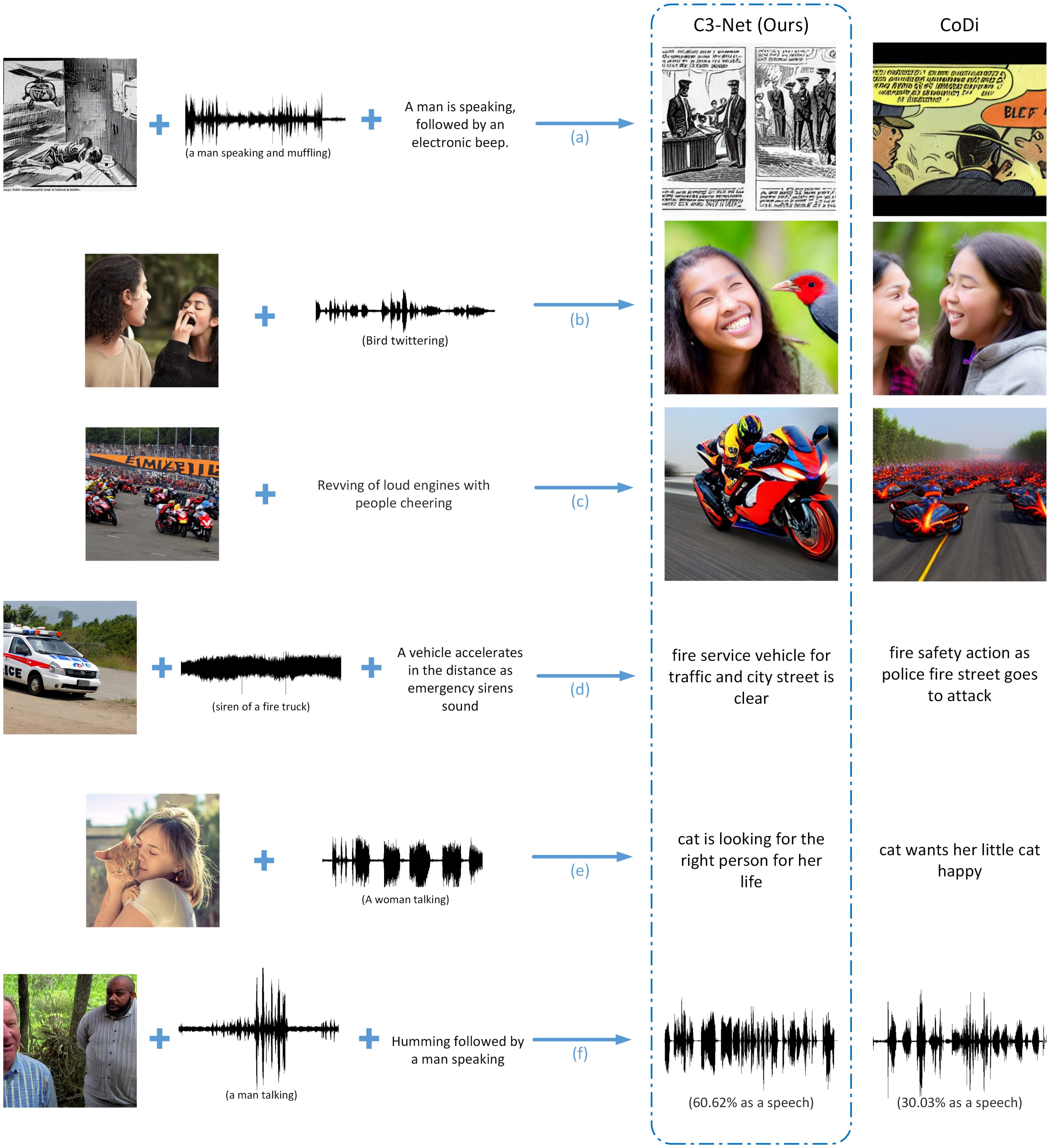
C3-Net (Ours)
CoDi
(a man speaking and muffling)
A man is speaking, followed by an electronic beep.
(a)
(Bird twittering)
(b)
Revving of loud engines with people cheering
(c)
(siren of a fire truck)
A vehicle accelerates in the distance as emergency sirens sound
(d)
fire service vehicle for traffic and city street is clear
fire safety action as police fire street goes to attack
(A woman talking)
(e)
cat is looking for the right person for her life
cat wants her little cat happy
(a man talking)
Humming followed by a man speaking
(f)
(60.62% as a speech)
(30.03% as a speech)

Figure 6. **Qualitative comparison** of compound-conditioned synthesis. The examples are conditioned on two or more images, texts, and audio conditions. *(a)* C3Net optimally extracts the feature of the blank-and-white sketch in the image condition. *(b)* C3Net better utilizes the audio condition, the sound of birds twittering. *(c)* C3Net generates an image of higher quality by focusing on the main subjects in the text condition. *(d)* The synthesized caption form C3Net has subject *fire service vehicle*, which is the optimal combination of subjects in all conditions. *(e)* C3Net synthesizes text including *person* and *cat*, where the baseline generation only has *cat*. This may be because the simple interpolation used in CoDi mixes the cat and person features into one subject. *(f)* The synthesized audio from C3Net rates higher in probability as a piece of speech classified by the model from [3].

aims to construct more meaningful latent modality structures. Addressing these challenges can improve the effectiveness of multimodal generative models, leading to more advanced and sophisticated content synthesis in the future.

References

[1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18208–18218, 2022. 1

[2] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023. 1

[3] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. BEATs: Audio pre-training with acoustic tokenizers. In *Proceedings of the 40th International Conference on Machine Learning*, pages 5178–5193. PMLR, 2023. 7, 8

[4] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023. 3, 5

[5] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7312–7322, 2023. 2

[6] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017. 6

[7] Yuan Gong, Cheng-I Lai, Yu-An Chung, and James Glass. Ssast: Self-supervised audio spectrogram transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10699–10709, 2022. 2, 5

[8] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988, 2022. 2

[9] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, 2021. 5, 6, 7

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 7

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851. Curran Associates, Inc., 2020. 2

[12] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2

[13] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 2

[14] Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In *International Conference on Machine Learning*, pages 13916–13932. PMLR, 2023. 2

[15] Qian Jiang, Changyou Chen, Han Zhao, Liqun Chen, Qing Ping, Son Dinh Tran, Yi Xu, Belinda Zeng, and Trishul Chilimbi. Understanding and constructing latent modality structures in multi-modal representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7661–7671, 2023. 7

[16] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet Audio Distance: A Reference-Free Metric for Evaluating Music Enhancement Algorithms. In *Interspeech*, pages 2350–2354, 2019. 7

[17] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *NAACL-HLT*, 2019. 6

[18] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014. 2

[19] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352*, 2022. 7

[20] Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven C.H. Hoi. Align and prompt: Video-and-language pre-training with entity prompts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 5

[21] Pengfei Li, Gang Liu, Jinlong He, Zixu Zhao, and Shenjun Zhong. Masked vision and language pre-training with unimodal and multimodal contrastive losses for medical visual question answering. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 374–383, 2023.

[22] Tao Liang, Guosheng Lin, Mingyang Wan, Tianrui Li, Guojun Ma, and Fengmao Lv. Expanding large pre-trained unimodal models with multimodal information injection for image-text multimodal classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15471–15480, 2022. 5

[23] Chin-Yew Lin. Rouge: a package for automatic evaluation of summaries. In *Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL*, pages 74–81, 2004. 7

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 6

[25] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. AudioLDM: Text-to-audio generation with latent diffusion models. In *International Conference on Machine Learning (ICML)*, 2023. 2

[26] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Improved image captioning via policy gradient optimization of spider. In *IEEE International Conference on Computer Vision (ICCV)*, pages 873–881, 2017. 6, 7

[27] Oscar Mañas, Pau Rodriguez, Saba Ahmadi, Aida Nematzadeh, Yash Goyal, and Aishwarya Agrawal. Mapl: Parameter-efficient adaptation of unimodal pre-trained models for vision-language few-shot prompting. *arXiv preprint arXiv:2210.07179*, 2023. 5

[28] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting on Association for Computational Linguistics*, page 311–318, 2002. 7

[29] Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Annual ACM Conference on Multimedia*, pages 1015–1018. ACM Press, 2015. 7

[30] Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. Automatic prompt optimization with "gradient descent" and beam search. In *EMNLP*, 2023. 1

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 3, 4, 5, 7

[32] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. 2

[33] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 3

[34] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1278–1286, 2014. 2

[35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2022. 1, 2, 5

[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2022. 6

[37] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 1

[38] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 6

[39] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *International Conference on Learning Representations (ICLR)*, 2023. 2

[40] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. In *Neural Information Processing Systems (NeurIPS)*, Red Hook, NY, USA, 2020. 2

[41] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations(ICLR)*, 2021. 2

[42] Yanan Sun, Zihan Zhong, Qi Fan, Chi-Keung Tang, and Yu-Wing Tai. Uniboost: Unsupervised unimodal pre-training for boosting zero-shot vision-language tasks. *arXiv preprint arXiv:2306.04715*, 2023. 5

[43] Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 3, 4, 5, 6

[44] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, 2015. 6, 7

[45] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. 6

[46] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 1, 2, 3, 5

[47] Ye Zhu, Yu Wu, Kyle Olszewski, Jian Ren, Sergey Tulyakov, and Yan Yan. Discrete contrastive diffusion for cross-modal music and image generation. In *International Conference on Learning Representations (ICLR)*, 2023. 1

[48] Heqing Zou, Meng Shen, Chen Chen, Yuchen Hu, Deepu Rajan, and Eng Siong Chng. UniS-MMC: Multimodal classification via unimodality-supervised multimodal contrastive learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 659–672, Toronto, Canada, 2023. Association for Computational Linguistics. 5