

# LANGUAGE-ORIENTED COMMUNICATION WITH SEMANTIC CODING AND KNOWLEDGE DISTILLATION FOR TEXT-TO-IMAGE GENERATION

<sup>†</sup>Hyelin Nam, <sup>‡</sup>Jihong Park, <sup>‡</sup>Jinho Choi, <sup>\*</sup>Mehdi Bennis, and <sup>†</sup>Seong-Lyun Kim

<sup>†</sup>Yonsei University, <sup>‡</sup>Deakin University, and <sup>\*</sup>University of Oulu

## ABSTRACT

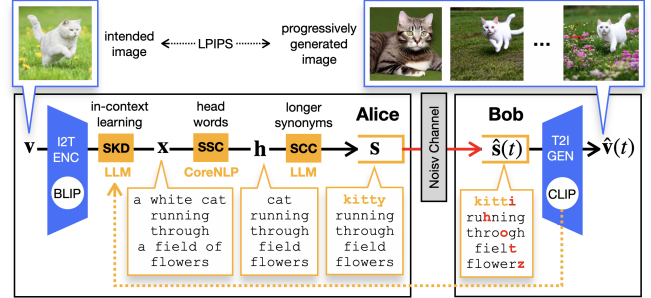
By integrating recent advances in large language models (LLMs) and generative models into the emerging semantic communication (SC) paradigm, in this article we put forward to a novel framework of language-oriented semantic communication (LSC). In LSC, machines communicate using human language messages that can be interpreted and manipulated via natural language processing (NLP) techniques for SC efficiency. To demonstrate LSC’s potential, we introduce three innovative algorithms: 1) semantic source coding (SSC) which compresses a text prompt into its key head words capturing the prompt’s syntactic essence while maintaining their appearance order to keep the prompt’s context; 2) semantic channel coding (SCC) that improves robustness against errors by substituting head words with their lengthier synonyms; and 3) semantic knowledge distillation (SKD) that produces listener-customized prompts via in-context learning the listener’s language style. In a communication task for progressive text-to-image generation, the proposed methods achieve higher perceptual similarities with fewer transmissions while enhancing robustness in noisy communication channels.

**Index Terms**— Semantic communication (SC), large language model (LLM), generative model.

## 1. INTRODUCTION

Semantic communication (SC) is an emerging research paradigm that focuses on the meanings (i.e., semantics) and effectiveness of communicating bits [1–5]. Deep joint source and channel coding (DeepJSCC) is a prime example wherein an encoder-decoder structured neural network (NN) acts as a transceiver, within which task-effective features are extracted from input data and made into communication messages. These *neural messages* are the NN’s hidden-layer activations trained and tailored for a specific task, which greatly improves communication efficiency [1–4].

This work was supported in part by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant (No. 2021-0-00347) funded by the Ministry of Science and ICT (MSIT), and in part by the Information Technology Research Center (ITRC) support program (IITP-2023-RS-2023-00259991) supervised by IITP. J. Park and S.-L. Kim are corresponding authors (email: jihong.park@deakin.edu.au, slkim@yonsei.ac.kr).



**Fig. 1.** An illustration of language-oriented semantic communication (LSC) for progressive text-to-image generation, empowered by semantic source coding (SSC), semantic channel coding (SCC), and semantic knowledge distillation (SKD).

However, neural messages constraint the full potential of SC. First, NN activations are not always universal messages, as they are influenced by their training data and communication environment. Indeed, separately trained transceivers in DeepJSCC are hardly interoperable without fine-tuning [6]. Furthermore, the semantics of these NN activations are nothing but what remains after achieving effective communication. It is therefore difficult to interpret and manipulate these semantics as intended.

By contrast, human language is universal and versatile to describe a broad range of tasks, owing to its evolution through-out diverse experiences in history. Moreover, with recent advances in natural language processing (NLP) and generative models, machines can interpret and manipulate human language. Motivated by this, in this paper we propose a novel *language-oriented SC (LSC)* framework, which facilitates SC through human *language messages*. The operation of LSC transceivers is trifold:

- 1) **Data-to-Language Translation:** Text-based cross-modal models transform input data into language messages to be transmitted (e.g., via CLIP for image-to-text (I2T) or Whisper for speech-to-text translation).
- 2) **Language Analysis & Manipulation:** Large language models (LLMs) and other NLP algorithms (e.g., GPT-4 [7], Llama 2 [8], and CoreNLP [9]) are utilized for analyzing

the syntax, semantics, and context in language messages and manipulating these messages for improving communication efficiency.

- 3) **Language-to-Data Generation:** Text-conditioned generative models produce intended data using the received message  $\mathbf{v}$  (e.g., via Stable Diffusion [10] for text-to-image (T2I) or Zeroscope [11] for text-to-video generation).

To showcase the potential of LSC, this paper we consider a point-to-point LSC scenario, where Alice sends a text prompt describing an intended image, while Bob progressively generates an image based on the accumulated received text prompt. The LSC accuracy is assessed using the learned perceptual image patch similarity (LPIPS) [12] that measures the distance between intended and generated images. To improve the communication efficiency of LSC, as visualized in Fig. 1, we focus on Step 2), and develop the following novel algorithms:

- **Head-based Semantic Source Coding (SSC)** is a lossy compression of the original prompt by pruning non-head words, inspired from our empirical findings that sending all words in a prompt does not always achieve the lowest LPIPS. The heads of a prompt are the words determining the syntactic category of the prompt, which can be identified, for example, using CoreNLP [9, 13].
- **Synonym-based Semantic Channel Coding (SCC)** adds redundancy into the prompt by replacing original head words with their longer synonyms, increasing the robustness to channel noise perturbing each character of the words. Only the synonyms ensuring the same semantics of the prompt are of interest, which can be found by, for instance, using GPT-4 [7].
- **In-Context Learning-based Semantic Knowledge Distillation (SKD)** aims to address the out-of-distribution (OOD) prompts due to different language knowledge between Alice and Bob, and enables Alice to emulate Bob’s prompts by assimilating Bob’s language knowledge. This can be achieved without re-training NN model parameters, by harnessing LLM’s unique capability of in-context learning, i.e., few-shot learning via demonstration [14].

Simulation results reveal SSC compresses transmitted messages by up to 42.6%, while surprisingly reducing LPIPS by 0.015 compared to full prompts. Applying SCC and SKD further cuts LPIPS by up to 0.007 and 0.009 by addressing channel noise and heterogeneous language knowledge.

**Related Works:** Recent research [3] employs generator models in SC but with neural messages, distinct from LSC’s language messages. The `ts.zip` [15] algorithm exploits LLM-based synonyms for compression, differing from our synonym-based SCC for robustness and from our head-based SSC. LSC stands apart from other language-based SC studies that mainly

focus on I2T compression or on the entropy of text truthfulness [16], in contrast to LSC harnessing LLM and NLP techniques to analyze and manipulate language messages for SC.

## 2. SEMANTIC SOURCE CODING FOR PROGRESSIVE TEXT-TO-IMAGE GENERATION

In this section, we propose SSC for a progressive text-to-image (T2I) generation task in a point-to-point communication scenario, as elaborated next.

- 1) **Image-to-Text Translation:** Alice has an intended image  $\mathbf{v}$  to send, and translates it into a text prompt  $\mathbf{x}$ , a sequence containing a set  $\mathbf{X}$  of words presented in a specific order, given as:

$$\mathbf{x} = \text{I2T}(\mathbf{v}) = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{|\mathbf{X}|}), \quad (1)$$

where  $\mathbf{x}_i$  is the  $i$ th word comprising  $|\mathbf{x}_i|$  characters. The function  $\text{I2T}(\cdot)$  represents an image-to-text (I2T) encoder such as BLIP [17] or CLIP [18].

- 2) **Head-based Semantic Source Coding (SSC):** Alice aims to compress and transmit text characters of  $\mathbf{x}$  while maintaining the semantics of  $\mathbf{x}$ . The semantics can be maintained when the key words of  $\mathbf{x}$  are presented without losing their syntax and context. To this end, SSC first identifies a set  $\mathbf{H}$  of  $\mathbf{x}$ ’s head words that determine the prompt’s syntactic category in linguistic analysis. While keeping head words’ order of appearance in  $\mathbf{x}$ , SSC produces a compressed sequence  $\mathbf{h} = (\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_{|\mathbf{X}|})$  in which:

$$\mathbf{h}_t = \begin{cases} \mathbf{x}_i, & \text{if } \mathbf{x}_i \in \mathbf{H} \\ \emptyset, & \text{otherwise.} \end{cases} \quad (2)$$

The head words in  $\mathbf{H}$  can be identified using the CoreNLP algorithm [13], i.e.,  $\mathbf{H} = \text{CoreNLP}(\mathbf{x})$ . Consequently, SSC yields the compression ratio  $|\mathbf{H}|/|\mathbf{X}| \leq 1$  in terms of words, and  $\sum_{\mathbf{x}_i \in \mathbf{H}} |\mathbf{x}_i| / \sum_{\mathbf{x}_i \in \mathbf{X}} |\mathbf{x}_i|$  in terms of characters.

- 3) **Text-to-Image Generation:** Bob receives the head words of  $\mathbf{h}_i$  in order, and progressively generates an image using a T2I generator such as Stable Diffusion [10] and DALL·E [19]. At the  $i$ -th head word reception with  $i \in \{1, 2, \dots, |\mathbf{H}|\}$ , the received prompt is  $\mathbf{h}(t)$ , and the generated image is:

$$\hat{\mathbf{v}}(t) = \text{T2I}(\mathbf{h}(t)) = \text{T2I}((\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_t)). \quad (3)$$

The perceptual similarity between Bob’s generated  $\hat{\mathbf{v}}(t)$  and Alice’s intended image  $\mathbf{v}$  is measured by the learned perceptual image patch similarity (LPIPS) score [12] that calculates the distance at hidden layers of pre-trained AlexNet, given as:

$$\text{LPIPS}(\mathbf{v}, \hat{\mathbf{v}}(t)) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \|f(\mathbf{v}) - f(\hat{\mathbf{v}}(t))\|_2^2. \quad (4)$$

The term  $l$  identifies the  $l$ -th layer having its width  $w$ , height  $h$ , and dimension  $H_l \times W_l$  with an activation function  $f(\cdot)$ .

**Table 1.** Message compression ratios of SSC with SCC and/or SKD.

Methods	w.o.SSC	SSC	SSC+SKD	SSC	SSC+SCC	SSC+SCC+SKD	SSC	SSC+SCC	SSC+SCC+SKD
	w.o. channel noise			SNR = 8.75 dB					
Compression ratio (Word)	1	<b>0.641</b>	0.654	<b>0.654</b>	<b>0.654</b>	<b>0.654</b>	<b>0.641</b>	<b>0.641</b>	<b>0.641</b>
Compression ratio (Character)	1	<b>0.426</b>	0.468	<b>0.468</b>	0.579	0.565	<b>0.426</b>	0.531	0.525
LPIPS	0.718	0.703	<b>0.697</b>	0.726	0.719	<b>0.715</b>	0.736	0.730	<b>0.721</b>

### 3. SEMANTIC CHANNEL CODING IN NOISY COMMUNICATION CHANNELS

In the previous section, we presume that Alice’s transmitted head words are perfectly received at Bob. In this section, we consider a noisy channel, and propose SCC to address noisy head word receptions at Bob.

1) **Noisy Channel Model:** Alice individually transmits a set  $C_{h_t}$  of characters in the head word  $h_t$ . Following a discrete memoryless channel (DMC) model, Bob receives the head word  $\hat{h}_t$  containing a set  $\hat{C}_{h_t}$  of characters, each of which is perturbed as a different character with a cross-over probability  $\epsilon > 0$  and is otherwise successfully received.

2) **Synonym-based Semantic Channel Coding (SCC):** In this noisy channel, SCC aims to enhance  $h_t$ ’s robustness by increasing  $|C_{h_t}|$  while maintaining the same semantics of the prompt  $h(t)$ . In the aforementioned channel, Bob encounters the same error in each characters following a geometric distribution. This does not allow communicating short words like “cat” that changes its semantics even with a single-character variation (e.g., bat, cut, and car), motivating SCC. In SCC, we consider a set  $S_{h_t}$  of candidate synonyms of  $h_t$ , given as:

$$S_{h_t} = \{s_1, s_2, \dots, s_{|S_{h_t}|}\}, \quad (5)$$

where  $s_t$  contains a set  $C_{s_t}$  of characters. Although  $S_{h_t}$  can be found using a dictionary, it ignores the context of  $h(t)$ , and does not guarantee the intended semantics. To solve this, SCC utilizes an LLM such as GPT-4 and Llama 2, a decoder-only autoregressive model that can predict the most in-context appropriate synonym  $s_t^*$  of  $h_t$  in  $h(t)$  by masking  $h_t$  and maximizing the following conditional unmasking probability:

$$s_t^* = \max_{s_j \in \mathcal{W}} p(s_j) = \Pr(s_j | h(t) \setminus h_t), \quad (6)$$

where  $\mathcal{W}$  is a set of total characters, e.g., 128 characters in ASCII. By relaxing this LLM, we obtain a set  $\hat{S}_{h_t}$  of in-context synonyms associated with their unmasking probabilities exceeding a threshold  $p_c > 0$ :

$$\hat{S}_{h_t} = \{\hat{s}_1, \hat{s}_2, \dots, \hat{s}_c\} = \{s_j | p_{s_j} \geq p_c\}. \quad (7)$$

Consequently, SSC can increase noise robustness of  $h_t$  within a set  $L_{h_t}$  of the levels in terms of characters, given as

$$L_{h_t} = \{|\hat{s}_j| \in \hat{S}_{h_t} | |h_t| \leq |\hat{s}_j| \leq L_c\}, \quad (8)$$

where  $L_c$  is the number of characters of the lengthiest synonym of  $h_t$ , i.e.,  $L_c = \lfloor \max_{\hat{s}_j \in \hat{S}_{h_t}} |C_{\hat{s}_j}| \rfloor$ .

### 4. SEMANTIC KNOWLEDGE DISTILLATION IN HETEROGENEOUS LANGUAGE KNOWLEDGE

In this section, we aim to address the problem when Alice and Bob have different knowledge on text-image relations by proposing SKD that enables Alice to produce Bob-customized text prompts via in-context learning.

1) **Heterogeneous Knowledge Model:** BLIP and CLIP are encoder-decoder NN models that store knowledge on image-text relations and text interpretations through cross-attention and self-attention weights. Suppose that Alice has BLIP for I2T while Bob utilizes CLIP encoder for word embedding in T2I. This incurs OOD generation in both, decreasing LPIPS.

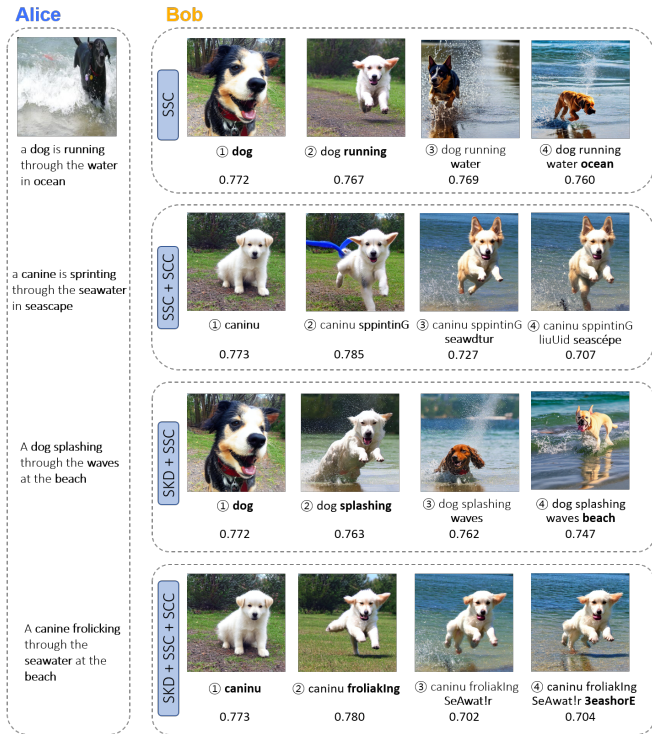
2) **In-Context Learning-based Semantic Knowledge Distillation (SKD):** A pre-trained LLM has excessive knowledge on spurious correlations, and conditioning it within a specific context can improve task performance. In-context learning enables this with few exemplary input-output pairs to teach the LLM a desired context, i.e., few-shot learning via demonstration [14]. Meanwhile, knowledge distillation (KD) is a method to transfer a target (teacher) model’s knowledge into another (student) by minimizing their output differences for common inputs [20]. Inspired from in-context learning and KD, SKD shows  $K$  input images to Alice and Bob that generate  $K$  output text prompts, using BLIP and CLIP, respectively, that are fed into an LLM for demonstration. This in-context learned LLM becomes a text-to-text (T2T) translator that can produce Bob-customized prompt  $\hat{x}_b$  for a given Alice’s prompt  $x_a$ :

$$\hat{x}_b = \text{T2T} \left( x_a; \{v^{(i)}, x_a^{(i)}, x_b^{(i)}\}_{i=1}^K \right), \quad (9)$$

where  $x_a^{(i)}$  and  $x_b^{(i)}$  are the  $i$ -th output prompts at Alice and Bob, respectively. After SKD,  $\hat{x}_b$  is fed into SSC and/or SCC. Note that SKD is applicable before SSC and/or after SCC. We focus on the former for simplicity.

## 5. NUMERICAL RESULTS

**Simulation Settings:** We consider that Alice’s I2T encoder is BLIP [21], and Bob’s T2I generator is Stable Diffusion v1.5 [10] that generates an image with 50 denoising steps. This diffusion process is conditioned by a text prompt encoded with CLIP encoder [18]. LLMs for SKD and SCC are based on GPT-4 [7], while SSC is run by CoreNLP [9, 13]. For image data, the Flickr8k dataset [22] is used, containing  $256 \times 256$

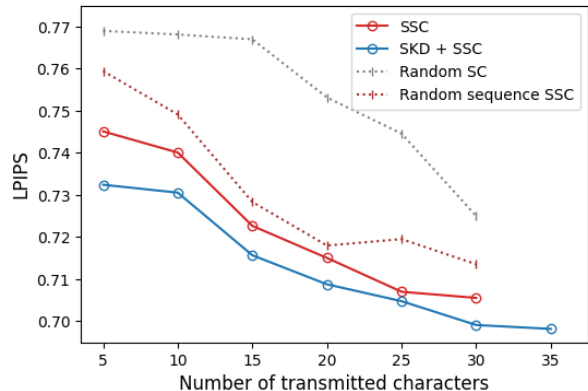


**Fig. 2.** Alice’s image and prompts (left) and Bob’s generated images and LPIPS (right), with SSC, SCC, and SKD.

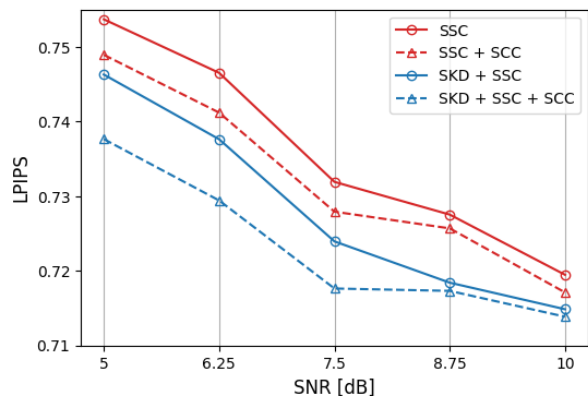
sized 8,092 samples. For sending prompts, each character is 8-bit ASCII coded, and modulated using 16QAM. During SCC,  $p_c$  is set as 0.72, and in SKD,  $K = 30$ . All LPIPS values are averaged over 100 simulation runs.

**Impact of SSC:** Tab. 1 reveals SSC achieves 64.1% word compression and 42.6% in character. Surprisingly, mean LPIPS improves by 0.015, suggesting SSC’s roles not only in compression but also in prompt engineering. Fig. 3 highlights SSC’s dual benefits (solid red), head extraction and appearance-based ordering. To dissect their LPIPS reduction contributions, we introduce two baselines: *Random SC* (dotted gray), which maintains appearance order with the same compression ratio as SSC but transmits random words from  $x$  instead of head words; and *Random sequence SSC* (dotted red), sending head words from  $h_t$  like SSC but in a shuffled order. Results indicate head extraction reduces mean LPIPS by up to 0.04, and appearance-based ordering contributes up to a 0.012 reduction, as seen when comparing SSC with Random SC and Random sequence SSC, respectively.

**Impact of SCC:** In Fig. 4, we observe that mean LPIPS decreases with SNR. Comparing SSC+SCC (dotted red) to SSC (solid red), SCC contributes to a reduction in mean LPIPS by up to 0.007. This reduction diminishes with SNR. However, SCC compromises compression, increasing it by up to 10.5% as shown in Tab. 1. In these simulations the increase in charac-



**Fig. 3.** SSC with or without SKD w.r.t. transmitted characters.



**Fig. 4.** SSC with or without SCC and SKD w.r.t. SNR.

ters is capped at 4. In certain instances, as illustrated in Fig. 2 at SNR = 7.5dB, the LPIPS reduction from SCC can be as much as 7.57 times its average. This suggests potential benefits in optimizing SSC level based on given channel conditions for future research.

**Impact of SKD:** As illustrated in Figs. 3 and 4, SKD contributes to a reduction in mean LPIPS by up to 0.006 and 0.009. Notably, the latter reduction surpasses even the contribution of SCC to LPIPS reduction. However, SKD may extend the prompt length, e.g., by an average of 5 characters in Fig. 3, highlighting a trade-off between compression and LPIPS.

## 6. CONCLUSION

In this article we proposed LSC, and developed SSC, SCC, and SKD that leverage NLP and LLM techniques to improve LSC’s SC efficiency under noisy channels and heterogeneous T2I/I2T knowledge. Future research might explore various tasks such as I2T-based control and compare LSC’s performance with its DeepJSCC counterpart.

## ACKNOWLEDGEMENT

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2021-0-00347, 6G Post-MAC), and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT).(No. 2023-11-1836)

## 7. REFERENCES

- [1] D. Gündüz, Z. Qin, I. E. Aguerri, H. S. Dhillon, Z. Yang, A. Yener, K. K. Wong, and C.-B. Chae, “Beyond transmitting bits: Context, semantics, and task-oriented communications,” *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 5–41, 2022.
- [2] Z. Qin, X. Tao, J. Lu, W. Tong, and G. Y. Li, “Semantic communications: Principles and challenges,” *arXiv preprint arXiv:2201.01389*, 2021.
- [3] S. Barbarossa, D. Communiello, E. Grassucci, F. Pezone, S. Sardellitti, and P. Di Lorenzo, “Semantic communications based on adaptive generative models and information bottleneck,” [Online]. *ArXiv preprint: arXiv:2309.02387*, 2023.
- [4] S. Seo, J. Park, S.-W. Ko, J. Choi, M. Bennis, and S.-L. Kim, “Towards semantic communication protocols: A probabilistic logic perspective,” *IEEE Journal on Selected Areas in Communications*, 2023.
- [5] H. Seo, J. Park, M. Bennis, and M. Debbah, “Semantics-native communication via contextual reasoning,” *IEEE Transactions on Cognitive Communications and Networking*, vol. 9, no. 3, pp. 604–617, 2023.
- [6] J. Choi, J. H. Park, S.-W. Ko, J. Choi, M. Bennis, and S.-L. Kim, “Semantics alignment via split learning for resilient multi-user semantic communication,” *submitted to IEEE Transactions on Vehicular Technology*, 2023.
- [7] OpenAI, “GPT-4 technical report,” 2023.
- [8] H. Touvron, L. Martin, K. Stone *et al.*, “Llama 2: Open foundation and fine-tuned chat models,” 2023.
- [9] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, “The Stanford CoreNLP natural language processing toolkit,” in *Association for Computational Linguistics (ACL) System Demonstrations*, 2014, pp. 55–60. [Online]. Available: <http://www.aclweb.org/anthology/P/P14/P14-5010>
- [10] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.
- [11] L. Khachatryan, A. Movsisyan, V. Tadevosyan, R. Henschel, Z. Wang, S. Navasardyan, and H. Shi, “Text2video-zero: Text-to-image diffusion models are zero-shot video generators,” 2023.
- [12] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.
- [13] D. Chen and C. D. Manning, “A fast and accurate dependency parser using neural networks,” in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 740–750.
- [14] T. Brown, B. Mann, N. Ryder *et al.*, “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [15] F. Bellard. `ts.zip`. [Online]. Available: [https://bellard.org/ts\\_server/ts.zip.html](https://bellard.org/ts_server/ts.zip.html)
- [16] R. Carnap, Y. Bar-Hillel *et al.*, “An outline of a theory of semantic information,” *Research Laboratory of Electronics, Massachusetts Institute of Technology*, 1952.
- [17] J. Li, D. Li, C. Xiong, and S. Hoi, “BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation,” in *International Conference on Machine Learning*. PMLR, 2022, pp. 12 888–12 900.
- [18] A. Radford, J. W. Kim, Hallacy *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.
- [19] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, “Zero-shot text-to-image generation,” 2021.
- [20] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” *NeurIPS Workshop on Deep Learning* [Online]. *ArXiv preprint: arXiv:1503.02531*, 2014.
- [21] J. Li, D. Li, C. Xiong, and S. Hoi, “BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation,” 2022. [Online]. Available: <https://arxiv.org/abs/2201.12086>
- [22] M. Hodosh, P. Young, and J. Hockenmaier, “Framing image description as a ranking task: Data, models and evaluation metrics,” *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–899, 2013.