Precision or Recall? An Analysis of Image Captions for Training Text-to-Image Generation Model

Sheng Cheng **Maitreya Patel** **Yezhou Yang**
Arizona State University
{scheng53, maitreya.patel, yz.yang}@asu.edu

Abstract

Despite advancements in text-to-image models, generating images that precisely align with textual descriptions remains challenging due to misalignment in training data. In this paper, we analyze the critical role of caption precision and recall in text-to-image model training. Our analysis of human-annotated captions shows that both precision and recall are important for text-image alignment, but precision has a more significant impact. Leveraging these insights, we utilize Large Vision Language Models to generate synthetic captions for training. Models trained with these synthetic captions show similar behavior to those trained on human-annotated captions, underscores the potential for synthetic data in text-to-image training.[1]

1 Introduction

Recent advancements in diffusion models such as Stable Diffusion (Rombach et al., 2022), DallE 3 (Betker et al., 2023), Emu (Dai et al., 2023), and Imagen (Saharia et al., 2022), have demonstrated remarkable capabilities in image synthesis. Despite these achievements, challenges persist in generating images that accurately align with the given text inputs (Huang et al., 2023). One issue is the misalignment between training captions and images, where captions either describe only a portion of the image or fail to describe the image content accurately. Recent research efforts have focused on enhancing caption quality using Large Language Models (LLM) (Fan et al., 2024) or Large Vision Language Models (LVLM) (Lai et al., 2023; Chen et al., 2024). Nevertheless, there is a scarcity of in-depth analysis on how specific factors influence the efficacy of text-to-image model training.

In this paper, we evaluate captions based on two metrics: precision and recall. We train the text-to-image (T2I) model using human-annotated captions that vary in levels of precision and recall. We then assess the model's capability for text-image alignment, specifically focusing on compositional ability. As illustrated in Figure 1, our findings indicate that while combinations of high precision and high recall yield the best results, generating captions with high precision is generally more beneficial. However, if the model is prone to hallucination, adding more diverse details can also enhance performance.

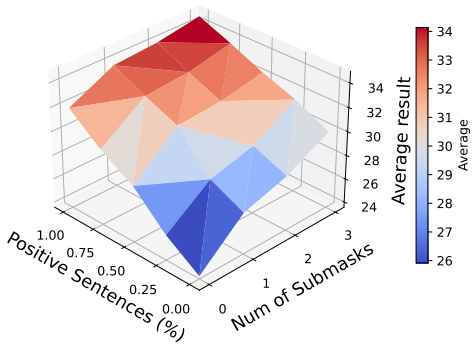

Figure 1: The result of the compositional capabilities across various combinations of precision and recall on human-annotated captions. Positive sentences indicate the precision of the captions, while the number of submasks represents the comprehensiveness of the captions.

Building on these insights as shown in Fgiure 1, we explore whether the same observation hold when using LVLMs to generate synthetic captions for T2I training. Given the variability across different LVLMs, some models may emphasize diversity at the expense of precision, leading to increased hallucination, while others may prioritize precision but sacrifice diversity. We evaluate the precision and recall of captions generated by various LVLMs and then use them to train T2I models. Our findings confirm that the compositional capabilities of the T2I models are consistent with our previous con-

[1]The data and code is available at https://github.com/shengcheng/Captions4T2I.

Findings of the Association for Computational Linguistics: EACL 2024, pages 3703–3709

clusions, underscoring the critical role of precision in caption generation.

The major contributions of this paper are:

- We systematically evaluate the impact of precision and recall on T2I model training, establishing that while both metrics are important, precision has a more significant influence on the model's performance.

- We extend our analysis by employing several LVLMs to generate synthetic captions. Our experiments show that the performance of T2I models trained with these synthetic captions is consistent with insights derived from human-annotated captions.

2 Related Work

Text-to-image diffusion model Given an input image x, it is paired with a corresponding caption c, which is segmented into multiple sentences $c_0, c_1, ..., c_N$. The sentence c_0 typically provides a general description of the image, while $c_1, ..., c_N$ detail the characteristics of specific subregions within the image. Current state-of-the-art text-to-image generation model is the latent diffusion model (Ho et al., 2020; Sohl-Dickstein et al., 2015; Nichol and Dhariwal, 2021; Rombach et al., 2022; Peebles and Xie, 2023). This model can be formulated as

$$\mathcal{L} := \mathbb{E}_{\varepsilon(x),c,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(c))\|_2^2\right],$$

where ϵ is the noise, t denotes the denoising timestep, θ represents the parameters of the diffusion model, ε and τ are image and text encoder.

Improving captions for text-to-image model training Models such as Pixart-α (Chen et al., 2024), DallE 3 (Betker et al., 2023), and Stable Diffusion 3 (Esser et al., 2024) emphasize the critical role of high-quality captions in their training. These systems incorporate captions synthesized by LVLMs (Wang et al., 2023; Liu et al., 2024; 202, 2023) into their training processes to enhance T2I generation capabilities. In particular, (Betker et al., 2023) investigates how synthetic captions contribute to improved T2I generation. However, these models do not specifically examine how caption quality impact the effectiveness of their training processes.

3 Analysis of Image captions for T2I training

Dataset Construction Our study utilizes the Dense Caption Dataset (Urbanek et al., 2023)[2], which comprises 8,012 images and 99,445 submasks derived from the SAM dataset (Kirillov et al., 2023). Each image in the dataset is segmented into multiple submasks, and both the main image and its corresponding submask images, known as subimages, are each paired with at least one detailed, human-annotated caption. We refer to the captions associated with the main image as the main captions and those linked to the subimages as subcaptions. These captions have been condensed to summarized versions of no more than 77 tokens using the LLAMA2-70B model (Touvron et al., 2023). To create negative captions, the LLAMA2-70B model modifies these sentences by altering their structure, editing content, or reshuffling words to form new sentences with the original vocabulary.

Our dataset is constructed based on these elements. In our study, we focus solely on how captions, including those for specific subregions within images, align with the overall image. Therefore, we do not use the subimages derived from submasks; instead, our dataset is built around the subcaptions associated with these submasks. To control the recall of the caption, each image is described not just by a basic caption that outlines the overall scene, but also by multiple subcaptions that provide detailed descriptions of specific regions, as shown in Table 1. We also manage the accuracy of our dataset by selectively including a certain proportion of negative captions or negative subcaptions.

# of submasks	0	1	2	3
Avg. nouns	12.3	20.5	28.0	35.2
Avg. tokens	54.8	91.0	125.1	157.8

Table 1: Data constructed analysis. The more submasks indicates more comprehensive description. The number of nouns is computed by Spacy (Honnibal and Montani, 2017). The number of tokens is computed by tokinzer (Raffel et al., 2020)

Training and Evaluation The family of Stable Diffusion models (v1.4, 1.5, and 2.1) (Rombach et al., 2022) integrates the CLIP (Radford et al.,

[2]CC-BY-NC license

Positive Sentences	Num of Submasks	Attribute Binding			Object Relationship		Average Result
		Color ↑	Shape ↑	Texture ↑	Spatial ↑	Non-Spatial ↑	
0%	0	0.1842	0.3079	0.3125	0.0884	0.2954	0.2377
0%	1	0.2322	0.3382	0.3389	0.1380	0.3031	0.2701
0%	2	0.2610	0.3340	0.3620	0.1446	0.3052	0.2814
0%	3	0.2870	0.3496	0.3840	0.1727	0.3066	0.3000
25%	0	0.2277	0.3042	0.3282	0.1230	0.2995	0.2565
25%	1	0.2801	0.3346	0.3620	0.1357	0.3047	0.2834
25%	2	0.2881	0.3356	0.3819	0.1529	0.3051	0.2927
25%	3	0.3203	0.3515	0.4193	0.1700	0.3078	0.3138
50%	0	0.2855	0.3271	0.3740	0.1243	0.3020	0.2826
50%	1	0.3250	0.3583	0.4180	0.1611	0.3062	0.3137
50%	2	0.3244	0.3529	0.4256	0.1727	0.3078	0.3167
50%	3	0.3358	0.3607	0.4291	0.1850	0.3082	0.3237
75%	0	0.3186	0.3402	0.3939	0.1350	0.3057	0.2987
75%	1	0.3387	0.3586	0.4443	0.1760	0.3078	0.3251
75%	2	0.3454	0.3679	0.4401	0.1784	0.3089	0.3281
75%	3	0.3581	0.3710	0.4503	0.1917	0.3095	0.3361
100%	0	0.3437	0.3599	0.4351	0.1507	0.3086	0.3196
100%	1	0.3500	0.3865	0.4695	0.1745	0.3091	0.3379
100%	2	0.3567	0.3872	0.4676	0.1832	0.3094	0.3408
100%	3	0.3717	0.3892	0.4662	0.1986	0.3100	0.3471

Table 2: The result of the compositional capabilities across various combinations of precision and recall on human-annotated captions. It contains five categories: color, shape, texture, spatial, and non-spatial. Positive sentences indicate the precision of the captions, ranging from 0% to 100%, while the number of submasks represents the comprehensiveness of the captions, ranging from 0 to 3.

2021) text encoder, which is limited to processing 77 tokens. To overcome this limitation, we employed the Pixart-α model (Chen et al., 2024), which utilizes the T5 (Raffel et al., 2020) text encoder capable of handling up to 512 tokens. We fine-tuned the model using LoRA (Hu et al., 2022) over 10 epochs, with batch size of 32 and learning rate of 1e-4. All experiments are run on an A100 GPU.

To validate the model's ability to generate images accurately aligned with the provided text, we use the T2I-Compbench (Huang et al., 2023), which includes five tasks. The first three tasks evaluate the model's capacity to accurately generate multiple objects in terms of correct color, shape, and texture. These tasks convert the text prompts into questions, which are tested using the BLIP2 model (Li et al., 2023). The spatial task examines the model's understanding of spatial directives like 'left' and 'right' using an object detection algorithm (Zhou et al., 2022). The non-spatial task focuses on object interactions and employs the CLIP model (Radford et al., 2021) to evaluate alignment.

Results Our results are shown in Table 2. We quantify the precision of the captions by the percentage of positive sentences and assess their recall through the number of submasks. It is important to clarify that a negative sentence does not necessarily indicate complete irrelevance to the associated image. Typically, such a sentence is mostly accurate but includes a few incorrect elements.

Our experiments confirm that both precision and recall influence the compositional capabilities of the model. However, our findings indicate a more significant impact of precision on performance as compared to recall. Notably, models trained with 0% positive sentences and three additional subcaptions underperform significantly relative to those trained with 100% positive sentences, even in the absence of any subcaptions, which contain approximately four times less information. At lower precision levels, increasing recall significantly boosts performance. For example, improving recall with captions that have 0% precision results in a 6.3% gain in performance. However, as the precision of captions improves, the benefits of increasing recall

Caption Method	T2I Model	Attribute Binding			Object Relationship		Average Result
		Color ↑	Shape ↑	Texture ↑	Spatial ↑	Non-Spatial ↑	
LLAVA	SD2.1	0.5406	0.4310	0.4941	0.1370	0.3165	**0.3838**
uform	SD2.1	0.5246	0.3878	0.5015	0.1250	0.3163	0.3710
BLIP	SD2.1	0.4962	0.4189	0.5087	0.1457	0.3156	0.3770
LLAVA	SDXL	0.4942	0.3942	0.4782	0.1452	0.3143	**0.3652**
uform	SDXL	0.4496	0.3458	0.4130	0.1185	0.3116	0.3277
BLIP	SDXL	0.4426	0.3652	0.4307	0.1395	0.3115	0.3379

Table 3: The result of the compositional capabilities on synthetic captions generated through different LVLMs.

decrease. When captions are 100% precise, the additional performance gain from increased recall is just 2.8%.

4 Insight for Synthetic captions for T2I training

Building on the findings from our previous analysis, this section investigates whether the observed impacts of precision and recall on performance also apply to synthetic captions.

We conducted experiments using synthetic captions generated by three different LVLMs: LLAVA (Liu et al., 2024), BLIP2 (Li et al., 2023), and uform[3]. Each model was given the instructions: *"Describe the image concisely,"* with a limit of fewer than 77 tokens per caption. The images for training are sourced from the MSCOCO dataset (Lin et al., 2014), which includes 118k images. We fine-tuned the Stable Diffusion v2.1 and Stable Diffusion XL base 1.0 (Podell et al., 2023)[4] model using these captions for 100,000 iterations with a batch size of 8 and a learning rate of 1e-4.

Due to the high cost of human verification for the precision and recall of synthetic captions, we use a modified version of the Faithscore (Jing et al., 2023) for evaluation. Initially, we filter out the descriptive content of the captions. Then, adapting from the original method in (Jing et al., 2023), which decomposes captions into ENTITY, COUNT, COLOR, RELATION, and OTHER, we refine our decomposition to better fit the compositional evaluation metric by using ENTITY, SHAPE, COLOR, TEXTURE, SPATIAL, NON-SPATIAL, and OTHER. Each sentence within these categories is then assessed for correctness. For the first two steps, we utilize the GPT-3.5 API (Brown et al., 2020), and for the final step of evaluation, we use the LLAVA1.5-13B model, following the approach detailed in (Jing et al., 2023). Due to the costs of the API usage, we randomly select 1000 samples to probe the quality of the captions. The result is presented in Table 4, including the number of entities per caption, which represents the recall of the captions, and their corresponding faithscore, which indicates the precision of the captions. It reveals that the BLIP model generates captions with less information but achieves high precision. Conversely, uform provides more diverse information but with relatively lower precision. Meanwhile, the LLAVA model not only maintains high precision but also exhibits better comprehensiveness compared to BLIP.

Method	LLAVA	BLIP	uform
# of Entites	4.90	2.12	**6.52**
Faithscore	0.911	**0.931**	0.831

Table 4: The number of entities and modified faithscore of synthetic captions generated from three different LVLMs model.

The results of the compositional capabilities of T2I models trained with synthetic captions are shown in Table 3. The findings reveal that the LLAVA model, which has relatively high precision and recall, outperforms the other two models. Despite containing three times less information than uform, the BLIP model's high precision enables it to perform better than the uform model. This observation aligns with insights from human-annotated captions, affirming that high precision is more crucial than high recall.

5 Conclusion

In this study, we investigated how caption quality affects T2I model training. We found that while

[3]https://huggingface.co/unum-cloud/uform-gen2-qwen-500m

[4]https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0

both precision and recall are important, precision is more crucial for effective training. These findings are confirmed using both human-annotated and synthetic captions from LVLMs. This insight could help improve the creation of synthetic captions for future T2I training.

Limitation

A key limitation of our study is the use of the LLAVA1.5-13B model in the Faithscore evaluation to determine the correctness of each entity in the image. Since synthetic captions are also generated with the LLAVA model, our evaluation might inherently favor captions generated by it. However, the LLAVA model remains one of the most advanced open-source Vision-Language Models available. Additionally, the cost of using human annotation for evaluation would be significantly high. In future work, we plan to explore using GPT-4 for evaluation to reduce this bias potentially.

Acknowledgements

This work was supported by NSF Robust Intelligence program grants #1750082 and #2132724. The authors acknowledge the resources provided by Research Computing at Arizona State University and the National Artificial Intelligence Research Resource (NAIRR pilot #240117). The authors also acknowledge technical access and support from ASU Enterprise Technology. The views and opinions of the authors expressed herein do not necessarily state or reflect those of the funding agencies and employers.

References

2023. Gpt-4v(ision) system card.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Junsong Chen, Jincheng YU, Chongjian GE, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. 2024. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *The Twelfth International Conference on Learning Representations*.

Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, et al. 2023. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*.

Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. 2024. Improving clip training with language rewrites. *Advances in Neural Information Processing Systems*, 36.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2023. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747.

Liqiang Jing, Ruosen Li, Yunmo Chen, Mengzhao Jia, and Xinya Du. 2023. Faithscore: Evaluating hallucinations in large vision-language models. *arXiv preprint arXiv:2311.01477*.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026.

Zhengfeng Lai, Haotian Zhang, Wentao Wu, Haoping Bai, Aleksei Timofeev, Xianzhi Du, Zhe Gan, Jiulong Shan, Chen-Nee Chuah, Yinfei Yang, et al. 2023. From scarcity to efficiency: Improving clip training via visual-enriched captions. *arXiv preprint arXiv:2310.07699*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR.

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

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

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jack Urbanek, Florian Bordes, Pietro Astolfi, Mary Williamson, Vasu Sharma, and Adriana Romero-Soriano. 2023. A picture is worth more than 77 text tokens: Evaluating clip-style models on dense captions. *arXiv preprint arXiv:2312.08578*.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. 2023. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. 2022. Simple multi-dataset detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7571–7580.

A Visualization of Human-annotated captions

We show sample that contain all positive sentences with 3 submasks (All-pos) and all negative sentences with 3 submasks (All-neg).

Figure 2: One sample from the DAC dataset, used for analysis of human-annotated captions.

All-pos: *A white double-decker bus and truck are parked at an intersection, with a urban skyline in the background. The bus has a Tissot logo and watch, while the truck has Chinese writing and phone numbers. The intersection has traffic lights, pedestrian crossing button, and yellow tactile paving for visually impaired people. A white*

Isuzu truck is parked on the right side, with a red 166PS logo on the roof and a driver visible through the right side door. There are four lines of Chinese words in red, blue, and green on the trunk, and three phone numbers at the bottom. Overcast sky with a darker gloom on the left. A white double-decker bus with a Tissot logo and watch picture under the windshield, an actress in red clothing, and a driver in blue uniform. The bus has 17 written on top and bottom, and people can be seen sitting and standing inside through the windows.

All-neg: *A sleek black double-decker bus and a rusty old truck are parked at a bustling roundabout, with a picturesque countryside landscape in the background. The bus features a large advertisement for a luxury fashion brand, while the truck is covered in colorful graffiti. The roundabout has a central fountain, lined with benches and surrounded by vibrant flower beds. A white Isuzu truck with a red logo on the roof has a driver visible through the right side door. The trunk features a colorful design with four lines of words, while three phone numbers are displayed at the bottom. The engine rumbles, indicating a powerful 166PS output. A blue sky with a hint of clouds. A white double-decker bus with a Tissot logo, an actress in yellow, and a driver in navy. The bus has 19 written on it and people are partying inside with streamers and balloons.*

B Visualization of Synthetic captions

We show sample that contain the synthetic captions from LLAVA, BLIP, and uform.

LLAVA: *The image features a tall clock tower with a blue and gold design. The tower is adorned with a cross on top, adding a religious touch to the structure. The clock is positioned towards the center of the tower, making it a prominent feature. The tower stands out against a blue sky.*

BLIP: *a clock tower on a building with a clock on top.*

uform: *A majestic brick clock tower with a blue and white tiled roof stands tall against a clear blue sky, featuring a cross at the top and a bell at the bottom. The tower is surrounded by other buildings, creating a serene urban landscape.*

Figure 3: One sample from MSCOCO, used for generating synthetic captions by LVLM.