ANCHOR: LLM-driven Subject Conditioning for Text-to-Image Synthesis

Aashish Anantha Ramakrishnan[1] **Sharon X. Huang**[2] **Dongwon Lee**[2]

[1]Optum AI [2]The Pennsylvania State University
[1]aashish.anantha.ramakrishnan@optum.com
[2]{suh972, dul13}@psu.edu

Abstract

Text-to-image (T2I) models have achieved remarkable progress in high-quality image synthesis, yet most benchmarks rely on simple, self-contained prompts, failing to capture the complexity of real-world captions. Human-written captions often involve multiple interacting subjects, rich contextual references, and abstractive phrasing, conditions under which current image-text encoders like CLIP struggle. To systematically study these deficiencies, we introduce ANCHOR, a large-scale dataset of 70K+ abstractive captions sourced from five major news media organizations. Analysis with ANCHOR reveals persistent failures in multi-subject understanding, context reasoning, and nuanced grounding. Motivated by these challenges, we propose *Subject-Aware Fine-tuning* (SAFE), which uses Large Language Models (LLMs) to extract key subjects and enhance their representation at the embedding-level. Experiments with contemporary models show that SAFE significantly improves image-caption consistency and human preference alignment, serving as a practical and scalable solution. The Dataset and code are available at: https://github.com/aashish2000/ANCHOR.

1 Introduction

Generative AI capabilities have grown rapidly, where large-scale pretraining has enabled powerful text-to-image (T2I) generation systems (Wang et al., 2023b). Yet, a fundamental challenge remains in aligning these systems with how language is naturally used in high-information settings. Most state-of-the-art T2I models are trained and evaluated on literal, object-centric prompts designed to directly reflect the visible contents of an image (Sharma et al., 2018; Chen et al., 2015). However, in many real-world contexts such as journalism, education, and social media, captions are often **abstractive**–i.e., they embed background knowledge,

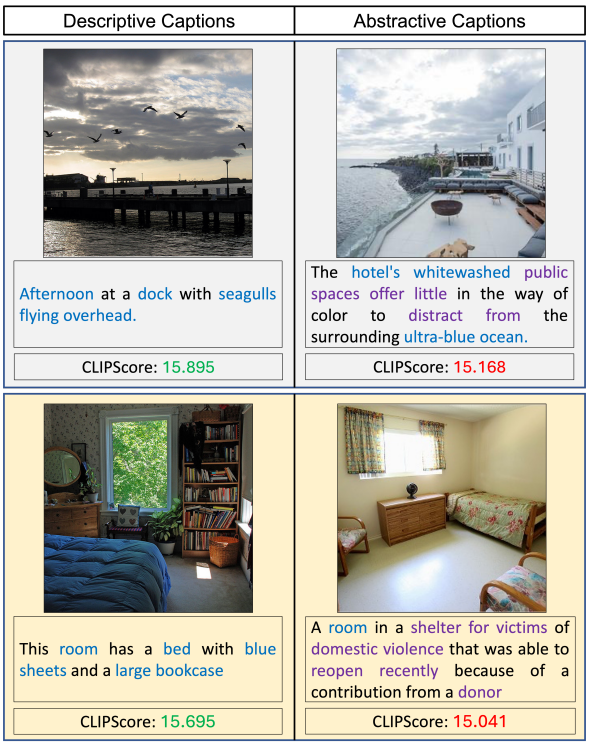


Figure 1: Example of descriptive prompts from the COCO Captions dataset (Chen et al., 2015) (Left) and abstractive captions from the ANCHOR (Right). Words highlighted in Blue directly translate to semantic objects while words highlighted in Purple signify contextual cues and syntactic variations, influencing the image indirectly. We also highlight abstractive captions differ at an embedding-level compared to descriptive prompts when paired with semantically similar images with lower CLIPScores.

reference multiple entities, and convey discourse-level intent (Grice, 1975; Alikhani et al., 2019; Vedantam et al., 2017).

In this paper, we position abstractive caption understanding as a critical yet underexplored challenge for T2I models and image-text encoders more broadly (Liao et al., 2024). While much progress has been made in aligning literal descriptions with visual content, real-world usage of Vision Lan-

Findings of the Association for Computational Linguistics: ACL 2026, pages 618–638
July 2-7, 2026

guage Models (VLM) require them to interpret language that leverages context information with implicit and narrative framings (Song and Zhou, 2021). With CLIP (Radford et al., 2021) being one of the most popular image-text encoders utilized for various VLMs (Schlarmann et al., 2024), our goal is to analyze how CLIP-based T2I pipelines handle this form of complexity and provide actionable insights for improvement. Existing studies isolate specific reasoning tasks such as multi-object scenarios (Abbasi et al., 2025) and spurious image-text correlations (Wang et al., 2024b) but aren't representative of real-world data domains. This is important as image-caption pairs may utilize multiple reasoning sub-tasks for intent comprehension.

To support this analysis, we construct ANCHOR, a large-scale dataset of over 70K abstractive image-caption pairs curated from news media. These captions reflect how humans naturally write about images in context: not just describing what is seen, but linking it to entities, events, and narrative discourse as depicted in Figure 1. The dataset is annotated to highlight four core challenges in real-world captioning: (1) *Semantic Objects*, (2) *Named Entities (NE)*, (3) *Contextual Cues*, and (4) *Syntactic Variations*.

Our empirical analysis reveals that current CLIP-based T2I models struggle with multi-subject understanding, context disambiguation and entity resolution in these settings. To better diagnose and address these issues, we propose *Subject-Aware FinE-tuning (*SAFE), a lightweight, modular strategy that augments T2I conditioning using **LLM-extracted subject guidance**. Rather than retraining encoders or introducing larger architectures, SAFE operates as a plug-and-play module that helps models differentiate between core semantic objects from contextual cues along with improved generalization on diverse caption styles. Thus, SAFE helps improve alignment without sacrificing scalability. Our primary goal is to use this framework as a lens to understand the specific failure modes of current systems when confronted with abstractive, multi-subject language. Our contributions are as follows:

- We introduce ANCHOR, the first large-scale dataset of abstractive, real-world captions for probing the limitations of current T2I models and image-text encoders.
- We develop SAFE, a subject-aware fine-tuning method that highlights the importance of disentangling semantic objects from contextual modifiers in abstractive captions.
- We provide a detailed evaluation of CLIP-based T2I models on ANCHOR, showing where and why they fall short—and how SAFE can serve as a targeted, interpretable remedy.

2 Related Work

Text-to-Image Synthesis There have been significant improvements in the field of T2I generation since the launch of Generative Adversarial Networks (GAN) (Goodfellow et al., 2014; Zhang et al., 2017; Xu et al., 2018; Zhu et al., 2019). As Transformer-based architectures (Vaswani et al., 2017) and model pre-training was successful on Natural Language Processing tasks, multi-modal encoders such as CLIP (Radford et al., 2021) significantly improved the quality of multi-modal embeddings and provided better input conditioning (Crowson et al., 2022; Zhou et al., 2022b). Diffusion models (Sohl-Dickstein et al., 2015) provided a breakthrough in training higher resolution models with greater expressivity by modeling generation as a reverse-Markov chain process (Nichol et al., 2022), (Ramesh et al., 2021; Ding et al., 2021). With the success of language model-based text-only encoders, Large Vision models adopt Large Language Model (LLM) based encoders for T2I generation, leveraging their text comprehension capabilities (Saharia et al., 2022). Flow matching-based models have also shown promise for data domain-agnostic generation tasks (Wang et al., 2025).

Multi-modal Reasoning Broadly, the types of reasoning tasks that models are commonly evaluated on are: Visual Reasoning, Context-based Reasoning, Factual Reasoning and Inter-modal Reasoning (Li et al., 2024a; Anantha Ramakrishnan et al., 2025a). With Visual Reasoning, MLLMs are assessed based on their ability to incorporate specific visual cues in a structured manner for tasks such as spatial and object relationship understanding (Thrush et al., 2022; Li et al., 2024b; Kamath et al., 2023; Zhang et al., 2025). The retrieval and interpretation of information from various sources related to specialized topics constitutes as Factual Reasoning (Lu et al., 2022; Wang et al., 2024a; Johnson et al., 2017). On the other hand, Context-based Reasoning explicitly measures how well models make use of the provided in-context

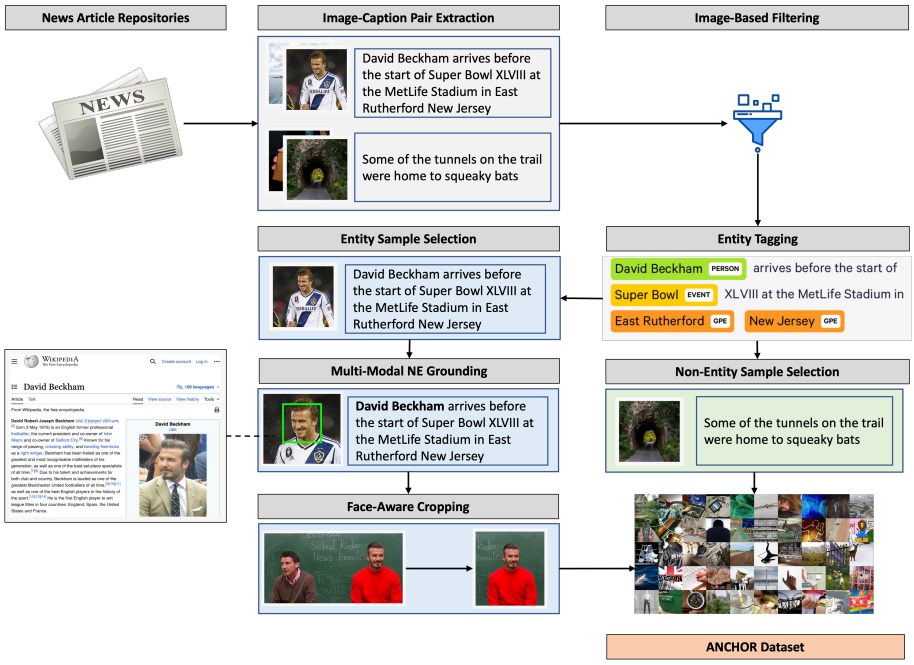


Figure 2: Overview of our dataset's pre-processing and filtering steps for both ANCHOR Non-Entity and ANCHOR Entity Subsets.

samples for tasks involving logical and compositional understanding (Schwenk et al., 2022; Zeng et al., 2024; Zong et al., 2024). Finally, Inter-modal reasoning involves interpreting causal linkages between modalities based on both semantic features, pragmatic cues and commonsense knowledge (Alikhani et al., 2020; Xu et al., 2022; Anantha Ramakrishnan et al., 2025b,c). Our proposed ANCHOR dataset evaluates both the context-based and inter-modal reasoning capabilities of T2I models using abstractive captions.

Datasets Initial benchmarks for T2I models focused on utilizing open domain images with short, descriptive text for evaluating their generation performance (Chen et al., 2015). In order to scale up diversity of evaluation prompts, image-text pairs crawled from online articles became more commonplace for facilitating both training and benchmarking these models (Sharma et al., 2018; Changpinyo et al., 2021; Schuhmann et al., 2021, 2022; Schuhmann C, Köpf A, Vencu R, Coombes T, and Beaumont R, 2022). With T2I models being pre-trained on larger and larger corpora, there has been a shift towards evaluation-only benchmarks with prompts to judge specific attributes of a generator's performance. PartiPrompts (Yu et al., 2022), DrawBench (Saharia et al., 2022) and UniBench (Li et al., 2022) provide diverse text prompts sorted based on style and difficulty. DiffusionDB (Wang et al., 2023c) is a large-scale collection of prompt-tuned caption-image pairs commonly used for sourcing captions for T2I evaluation. All these benchmarks focus on captions that provide accurate descriptions of physical objects within images. We aim to include captions containing situational context information and complex sentence structures as a part of ANCHOR.

3 ANCHOR: Dataset Construction

The ANCHOR (Abstractive News Captions with High-level cOntext Representation) dataset is a large-scale image-caption pair dataset extracted from news articles as shown in Figure 2. To construct this, we use open-source news image captioning datasets: VisualNews (Liu et al., 2021) and NYTimes800K (Tran et al., 2020). For effectively testing caption comprehension of image-text encoders, we need to isolate the impact of caption structures from other factors that influence the synthesized image quality. NE features such as faces of specific people can pose a significant challenge to generators (Rombach et al., 2022), (Ramesh et al., 2022). This is likely due to the complexity of learning NE features compared to more generic visual concepts during pre-training. To assess if artifacts generated by these models are due to poor understanding of implicit context or specific entity features, we split our data into

2 distinct subsets: ANCHOR Non-Entity & ANCHOR Entity. Similar to other internet-sourced datasets (Sharma et al., 2018), we remove 95% of low-quality image-caption pairs from a combined 1.8M samples while pre-processing as described in Appendix Section A.

3.1 ANCHOR Non-Entity Subset

This subset contains 72692 samples selected from articles published by 5 different news media organizations. The train / val / test split of the dataset is in the ratio 90% / 5% / 5% respectively. We ensure that the associated images do not contain representations of NEs in order to evaluate the influence of different caption components independently. Using a RetinaFace-based (Deng et al., 2020) face detector, we flag images containing identifiable faces.

3.2 ANCHOR Entity Subset

With NEs being a critical component of news image-caption pairs, this subset has been designed to evaluate the impact on T2I reasoning when NEs are present along with the other identified linguistic structures. This subset contains 7516 image-caption pairs with 48 different NEs. The current subset primarily includes PERSON entities. This is due to their frequency of mentions in news media, and consistency of physical features across images. Since there is a long-tail distribution of images per NE, we construct an eval set by selecting only 50 image-caption pairs per NE for our experiments.

Multi-Modal NE Grounding The challenge with NE mention detection is that entities can be referred to by different names according to the situation. Example: David Beckham can be referred to as: "Beckham", "David", "David Robert Joseph Beckham", etc. To avoid this ambiguity, we need to reliably link each mention to a real-world entity. We perform Multi-modal NE grounding to link each entity mention using Wikipedia as a real-world knowledge source. Using the REL Entity Linker (van Hulst et al., 2020), we extract entity mentions from the previously selected samples and link them to their appropriate Wikipedia pages. We used a Wikipedia dump from 2019-07 as our knowledge source. Although this helped in removing erroneous mentions detected from text captions, we also need to ensure each image contains the mentioned entity. Using their linked Wikipedia pages, we download the main image and create a repository of reference images for each entity in our dataset. Since our focus is on PERSON entities, we ensure that a face is detected in each of the downloaded reference images. A FaceNet-based (Schroff et al., 2015) face recognition module is used to ground each image to an entity category.

Face-Aware Cropping The non-centered nature of foreground objects in images poses a challenge for consistent image evaluation. Many photographs are taken as long shots with the entity's face in different sections of the image. To standardize these images, we crop and resize the images taking into account the target entity position. By extracting the bounding boxes of our target entity face, we calculate its centroid as a reference coordinate for cropping. We then take a fixed window crop of the entity image such that the entity centroid is aligned closely to the center of the crop. This approach of Face-aware cropping helps maximize the image area occupied by an entity and further isolates its physical features. Through this process, we can evaluate the visual features of different entities.

3.3 Dataset Quality and Diversity

We show how lexically and semantically diverse the captions in ANCHOR are compared to other popular datasets through quantitative metrics such as token diversity and CLIPScore in Appendix Section B. To further assess the overall quality of the dataset and the number of image-caption pairs that are abstractive, we conducted a human evaluation study on Amazon MTurk tabulated in Table 1. We consider a random sample of 300 samples extracted from the test split of ANCHOR. Out of the 300 selected images, 200 belong to the non-entity subset and 100 belong to the entity subset. Using this extracted sample, we perform a human evaluation of our dataset quality. The two questions we mainly aim to answer through this evaluation are: (1) Are the image-caption pairs closely related to each other from a human perspective? and (2) Are these captions Abstractive in nature? We launched our survey with 150 unique participants and each participant rated 10 samples. Per Image-caption pair, we collect 5 responses amounting to a total of 1500 responses.

We observe that 97% of surveyed samples are related to each other. We also see that 89.3% of the related captions are rated as Abstractive in nature containing atleast one of the identified caption components. This supports our hypothesis and validates that our dataset pre-processing pipeline produces

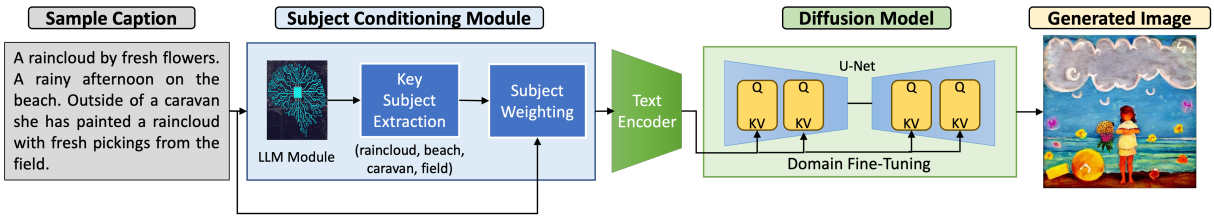


Figure 3: Overview of our Subject-Aware FinE-tuning Approach (SAFE). Through the Subject Conditioning Module and Domain Fine-tuning, we enable better comprehension of context information in abstractive captions.

Name	Size	Related	Abstractive	Descriptive
ANCHOR Full	300	291	**260**	31
ANCHOR Entity	100	99	**93**	6
ANCHOR Non-Entity	200	192	**167**	25

Table 1: Human Evaluation of ANCHOR Dataset Quality. Our studies show that ANCHOR contains highly related captions with a majority of them being abstractive in nature.

high-quality image-caption pairs.

4 Understanding Abstractive Captions with External Knowledge

At the most basic level, a subject is defined as an individual element or a concept that provides either visual, situational or complementary meaning within a caption (Cohn, 2003). We categorize the components of an abstractive caption into 4 main types: *Semantic Objects, Contextual Cues, NEs, and Syntactic Variations.* Subjects such as noun phrases that have commonly identifiable visual patterns are defined as *semantic objects*. *Contextual cues* inform readers about the situational information surrounding a particular caption such as the setting of the image, the events leading up to that moment, and why the image's contents are significant (Song and Zhou, 2021). *NEs* are domain-specific terms that refer to real-world objects and their specific features. Commonly employed NE categories include names of people, places, organizations, etc. Finally, *syntactic variation* corresponds to the different forms of expression that are achieved by manipulating the relationships between different visual objects (Bugliarello and Elliott, 2021). It helps in the comprehension of complex sentences with multiple subjects and connecting clauses.

Descriptive prompts primarily utilize semantic objects with minimal presence of syntactic variations, contextual cues, and NEs while abstractive captions from domains such as news media take full advantage of all these features (Zhou et al., 2022a; Anantha Ramakrishnan et al., 2025c). For the task of T2I generation, we utilize conditioning to conform the generated image to certain criteria provided in the text input. Text conditioning in T2I models is accomplished through embeddings extracted by text encoders such as CLIP. Encoders typically employ self-attention to analyze and assign importance scores to individual elements of the input sequence while retaining the global context information across the entire sequence. This simplifies the task of word importance estimation during the generation process with every word contributing directly towards a visual concept. Let S_{desc} correspond to a descriptive caption, we can represent it as a collection of subject tokens T_i where $i = 1, \ldots, m; m \in Z$.

$$S_{desc} = \{T_1, T_2, \ldots, T_i, \ldots, T_m\} \quad (1)$$

Every subject is expected to either define or describe the properties of a visual concept present in the generated image. When generating embeddings E_{desc} for a caption, each subject token is accompanied by a weight co-efficient α_i.

$$\begin{aligned} E_{desc} &= TextEnc(S_{desc}) \\ &= \Sigma_{i=1}^{m} \alpha_i * T_i \end{aligned} \quad (2)$$

In the case of abstractive captions S_{abstr}, the goal of a T2I generator is to identify the salient subjects describing image contents while incorporating context information selectively. This becomes a particularly challenging task when there are multiple subjects present in the caption. Certain subjects may get forgotten or misrepresented in the generated image. This problem of "Prompt Following" (Betker et al.) has been documented with real-world captions. To boost comprehension of T2I generators, we aim to explicitly modify the weights α_i for subject tokens T_i. Since current encoders can capture all the relevant subjects from an image caption, we attempt to increase the weights

Model	FID_{CLIP} (↓)	IR (↑)	HPS V2 (↑)
SAFE (DFE + GPT-3.5)	<u>7.2804</u>	**0.0664**	**0.2393**
Stable Diffusion 2.1 (CR)	10.6595	-0.3388	0.2101
Stable Diffusion 2.1 (LoRA)	**6.9222**	-0.0861	0.2329
Stable Diffusion 2.1 (Base)	7.4780	<u>0.0251</u>	<u>0.2385</u>
Stable Diffusion 1.5 (Base)	7.4742	-0.0925	0.2188

Table 2: Results of Abstractive Text-to-Image synthesis on ANCHOR Non-Entity Subset. Images generated using SAFE show higher image-caption alignment while not sacrificing image fidelity compared other baselines.

of key subjects T_{key} describing the main components of an image. Let W_{abstr} correspond to the vector containing the scale multiplier $\beta > 1$ for each token. The embeddings for abstractive captions E_{abstr} will be conditioned using the scale multipliers found in W_{abstr}. This scale multiplier helps align embeddings toward the intended meaning of a caption by acting as a prompt weight.

$$W_{abstr} = \begin{cases} \beta, & \text{if } T_i \in T_{key} \\ 1, & \text{otherwise.} \end{cases} \quad (3)$$

$$E_{abstr} = TextEnc(S_{abstr}) \odot W_{abstr} \quad (4)$$

4.1 Subject-Aware FinE-tuning (SAFE)

As illustrated in Figure 3, SAFE utilizes a Stable Diffusion-based backbone for T2I generation. By taking advantage of prompt weighting and fine-tuning strategies, our approach aims to enhance both the image fidelity and prompt following capabilities over baselines.

LLMs for Subject Conditioning The key challenge in implementing subject conditioning is identifying which subjects/phrases to weigh positively. To replace human guidance in the process of prompt weighting, we evaluate the use of LLMs in identifying salient subjects from sentences. Leveraging the commonsense reasoning abilities of LLMs, we utilize instruction-based prompting to extract salient subjects from each sentence. Subject identification is done in a zero-shot manner using only the prompt and the pre-trained world knowledge of LLMs. This process allows us to explicitly condition the input embeddings in an observable and explainable manner. Compared to other LLM-based grounding strategies (Lian et al., 2024; Feng et al., 2023), subject conditioning requires only single-stage prompting and also utilizes fewer tokens per generation. Additionally, we also compare different LLM architectures including both Commercial (GPT-4, GPT-3.5) and Open-source (Mixtral 8x7B, Orca Mini-13B) (Jiang et al., 2024; Mukherjee et al., 2023) models on this task as described in Appendix Section C. This comparative study helps us ascertain the types of models capable of producing high-quality subject weights.

Handling Domain-Shift Real-world images and captions have specific characteristics that differ from the general outputs generated by T2I models. Specifically, real-life photographs with specific foreground and background objects are present in abundance compared to artistic or animated-style images. To tackle this domain shift, we develop our Domain fine-tuning (DFE) strategy on ANCHOR. Traditional Mean Squared Error-based losses used for Stable Diffusion fine-tuning tend to generate unrealistic images (Zhang et al., 2018; Lin and Yang, 2023), making them unsuitable for our task. We adopt the Reward Feedback Learning (ReFL) (Xu et al., 2023) strategy for directly optimizing Stable Diffusion on a reward model trained to score image-caption alignment. The selected reward model ImageReward (IR) (Xu et al., 2023) has been trained on 137K human-annotated samples to predict alignment between image-caption pairs. Our proposed improvements in DFE over vanilla ReFL focus on improving both alignments with the ground truth image and caption, instead of only caption alignment. In DFE, we initialize the latent vector of Stable Diffusion based on the ground truth images for each caption instead of random initialization as implemented in ReFL. This helps control the diffusion process in generating target distribution-aligned images. To increase training stability, we fine-tune only the Attention Layers using Low Rank Adaptation (LoRA) (Hu et al., 2022). Additionally, the noise scheduler of the original ReFL pipeline was limited to having 40 timesteps. The authors identified that latents sampled between timesteps 30-39

Original	SD 2.1 (Base)	SD 2.1 (CR)	SAFE (DFE + GPT-3.5)
	 (Missing Shoes)	 (Misleading Style)	 (Shows shoes & clothes)

Ex1: The school offers clothing, including shoes, to its students.

	 (Missing Case)	 (Missing Keyboard)	 (Shows case with keyboard)

Ex2: The Galaxy Note 5 can be used with a case that doubles as a physical Qwerty keyboard to aid typing.

	 (Incorrect Human Subject)	 (Incorrect Context)	 (Shows correct context)

Ex3: A Faraday bag, which blocks remote signals to devices such as cellphones and tablets.

Figure 4: Qualitative comparison of different T2I models on ANCHOR Non-Entity. Words highlighted in Orange are used for subject weighting in SAFE. Incorrect Visual Artifacts in images are described in Red and the correct contextual expectations for generated images are marked in Green.

produced distinguishable IR scores. We extend this by our noise scheduler to 100 timesteps and sampling from steps 40-99 for loss calculation.

5 Experiments and Results

5.1 Evaluation Metrics

To holistically evaluate the samples generated by T2I models on ANCHOR Non-Entity, we report 3 different types of metrics: CLIP-based Frechet Inception Distance (FID_{CLIP}) (Kynkäänniemi et al., 2022), IR (Xu et al., 2023), and Human Preference Score (HPS) V2 (Wu et al., 2023). On ANCHOR Entity, we include Identity Preservation and Face Detection accuracy as additional metrics to quantify their entity image generation performance similar to (Yuan et al., 2023). We provide additional details on metric selection and model hyperparameters in Appendix Section D.

5.2 Baseline Models

Our goal is to study *encoder-level* behavior in a controlled setting, thus we compare single-encoder models primarily as our baselines. Newer architectures that rely on multiple text encoders, LLM-based encoders, and adapter/fusion layers mix several embedding streams, making encoder-level analysis difficult (Toker et al., 2024; Gao et al., 2024). Thus we select: Stable Diffusion V1.5 and V2.1 (Rombach et al., 2022) as representative models for T2I generation. These models utilize CLIP-based architectures for image-text encoders, showing strong performance in traditional benchmarks such as COCO Captions (Chen et al., 2015). Additionally, we also compare the performance of LLM-powered Caption Rewriting (CR) for translating abstractive captions into descriptive text. Specifically, we use an LLM to rewrite a caption into

an instruction prompt of the format "Generate an image ...".

5.3 Quantitative Results

On analysis of the scores presented in Table 2, SAFE outperforms both baseline Diffusion models on the ANCHOR Non-entity test set. Similarly, LoRA fine-tuning only improves on image fidelity, while showing lower image-caption alignment. The explicit guidance through both fine-tuning and subject conditioning contributes towards capturing the intended meaning of captions and also producing more aligned images without changing the encoder architecture. Caption rewriting on the other hand performs significantly worse on our benchmark metrics, signaling that CLIP-based encoders face challenges in comprehending situated context even when translated descriptively.

For ANCHOR Entity, we report the average metric scores across all entity classes in Appendix Table 8. Our results show that captions with NEs are highly sensitive to visual degradation upon embedding re-weighting (Xiao et al., 2025), as evidenced by the overall lower Detect and Identity scores. Although multiple entities may be better represented through re-weighting, the overall quality of the generated image is impacted adversely, demonstrating the challenging nature of this problem. This justifies our development of two subsets within ANCHOR to support this analysis.

5.4 Qualitative Results

From the examples presented in Figure 4, we observe that SAFE achieves more faithful prompt interpretation than the baselines by effectively handling contextual cues and syntactic variations within captions. In the first example, the subjects "clothing", "shoes", and "students" are semantically related but require correct syntactic parsing to preserve their relationships. SAFE successfully captures these dependencies, producing images with visible shoes and clothing. In contrast, SD 2.1 (Base) fails to represent all key subjects, particularly the shoes. SD 2.1 (CR) generates a more grounded image but misinterprets the contextual framing, rendering the scene as a sketch rather than a realistic photo. This reflects how caption rewriting can distort the original style cues embedded in the caption structure. In the second example, a contextual focus is placed on the "keyboard", requiring the model to understand both the object and its role within the scene. SD 2.1 (Base) again overlooks this nuance, producing an image where the keyboard is missing. SAFE, however, correctly maintains the contextual intent and learns the syntactic variation of the caption, resulting in a composition that aligns closely with the described scenario. Human evaluation results in Table 3 further confirm that images generated by SAFE are consistently preferred for their stronger contextual grounding and structural coherence. Details regarding study setup and design are provided in Appendix E.1.

Model	Total	Preferred Samples	Preference (%)
SAFE (DFE + GPT-3.5)	185	102	**55.13**
Stable Diffusion 2.1	185	83	44.86

Table 3: Human Evaluation Results on ANCHOR. We show that there exists higher preference among users for images generated by SAFE compared to T2I baselines.

Model	FID_{CLIP} (↓)	IR (↑)	HPS V2 (↑)
SAFE (GPT-3.5)	<u>7.2614</u>	0.0624	<u>0.2392</u>
SAFE (GPT-4)	**7.2482**	<u>0.0673</u>	0.2391
SAFE (Mixtral 8x7B)	7.2649	**0.0723**	**0.2394**
SAFE (Orca Mini-13B)	7.3571	0.0298	0.2381

Table 4: Ablation study evaluating the impact of different LLM Agents for subject weight extraction. Open Source LLM backbones perform equally well when compared to proprietary models for subject weighting.

On ANCHOR's Entity Subset, the generated images using SAFE cohesively include most objects presented in the caption, while the images generated by SD 2.1 (Base) focus primarily on the NEs present. As shown in the Appendix Figure 6, although SD 2.1 (Base) does have challenges such as the repeatedly generating the same entity, the overall quality of NE features is higher. This presents itself as better alignment across metrics.

5.5 Ablation Study

Subject Weight Quality Across LLM Models To assess the variation in commonsense reasoning and world knowledge of different LLM architectures, we collect subject weights from 4 different LLMs: GPT-3.5, GPT-4, Orca Mini-13B, and Mixtral 8x7B Mixture of Experts (MoE). For our ablation study, we replace only the provided subject weights from each model during inference using SD 2.1 (Base) as our T2I backbone as shown in Table 4. We can observe a clear correlation between LLM performance on other commonsense reasoning tasks and key subject delineation with Mixtral and GPT-4 outperforming other models. Subject

weights generated by models with lower number of parameters like Orca Mini-13B still show improvements on 2 of the 3 metrics compared to our baselines. This demonstrates the potential of open-source LLMs in boosting caption understanding for cross-modal generative tasks.

Model	FID_{CLIP} (↓)	IR (↑)	HPS V2 (↑)
SAFE (DFE + GPT 3.5) (x^2)	<u>7.2804</u>	**0.0664**	<u>0.2393</u>
w/o GPT-3.5	7.4851	0.0249	0.2385
w/o DFE	**7.2614**	<u>0.0624</u>	0.2392
SAFE (DFE + GPT 3.5) (x^1)	7.3729	0.0564	**0.2395**
SAFE (DFE + GPT 3.5) (x^3)	7.3049	0.0040	0.2361
SAFE (DFE + GPT 3.5) (x^4)	7.8825	-0.1835	0.2255

Table 5: Ablation study evaluating the effectiveness of different scale multiplier values (x^n) and model components.

Impact of Subject Conditioning We investigate the impact of each of SAFE's components on generation quality as shown in Table 5. We observe that Subject Conditioning provides a significant contribution towards the observed metric performance. The addition of DFE also boosts image-caption understanding without majorly impacting image fidelity, as reflected in all 3 metrics presented. The positive correlation between IR and HPS V2 even with the addition of DFE, confirms that the fine-tuning process hasn't overfit on the reward model.

Impact of Subject Scale Multiplier We evaluate various candidate score multipliers as shown in Table 5. Here, x^1 refers to a scale factor of 1.1, x^2 refers to a scale factor of $(1.1)^2$, and so forth. We selected a multiplier of x^2 as it scores the highest in 2 out of 3 metrics tested. Increasing it beyond x^2 does not provide any meaningful improvements.

Generalizability of Subject Weighting across T2I Benchmarks To understand if our proposed subject weighting methodology is generalizable to other task domains apart from news image synthesis, we evaluate our approach on existing benchmark T2I datasets. We selected the Conceptual Captions (CC3M) dataset (Sharma et al., 2018) as a relevant benchmark. Since CC3M is sourced from web-scraped articles and blog posts, it offers a comparable level of caption complexity to ANCHOR. Our experiments cover both baseline and SAFE models on the public validation set of CC3M as described in Table 6. We show that our proposed weighting approach improves context relatedness and human preference alignment, even without any domain fine-tuning.

Model	FID_{CLIP} (↓)	ImageReward (↑)	HPS V2 (↑)
SAFE (GPT 3.5)	8.4124	**0.2700**	**0.2521**
Stable Diffusion 2.1	**8.3883**	0.2414	0.2516

Table 6: Ablation Results of T2I synthesis on CC3M dataset. We show the extendability of SAFE to other T2I benchmarks that contain Non-Entity captions.

6 Conclusions

We present ANCHOR, a novel dataset for evaluating image-text encoders such as CLIP on abstractive captions, identifying key challenges in multi-subject understanding and context-based reasoning. To mitigate this, our subject conditioning strategy: SAFE helps improve subject grounding and interpreting syntax variations. With SAFE being able to re-rank the importance of specific subjects at an embedding-level, we improve contextual alignment without increasing parameter size or retraining the encoder. Through fine-tuning, we integrate both image-level and human-preference alignment objectives, building on top of traditional techniques such as LoRA and ReFL. Compared to other LLM + Diffusion methods (Liao et al., 2024; Lian et al., 2024), SAFE requires only one LLM query with significantly fewer tokens returned, lowering inference cost and processing time.

Limitations

Since we build on top of open-source Large Foundational Models such as Stable Diffusion, GPT-3.5, GPT-4, Mixtral-8x7B, and Orca, our approach inherits all their biases. We do not analyze T2I models that use multi-encoder architectures (Esser et al., 2024; Chen et al., 2023) given the challenges in disentangling the influence of individual encoders towards multi-subject and context comprehension. Extending SAFE to multilingual abstractive captions also faces several challenges in validating the abstractiveness of publicly available datasets and the inherent weakness of multilingual CLIP variants in performing contextual grounding on low-resource languages (Ananthram et al., 2024). The lack of task-specific fine-tuning to improve entity likeness generation is another limitation that our approach faces. Future research directions include development of entity concept datasets and analyzing unified multi-modal and multi-lingual

encoders that learn on both image and text tokens simultaneously.

Ethical Considerations

We only source data from publicly available news media repositories that are licensed for use in research. We will also release our dataset under the same license restrictions for public access (CC BY-NC-SA 4.0). To remove NSFW content, we apply a combination of profanity word-lists (Nguyen) and machine learning-based caption filtering methods (Zhou). Our goal for launching a news media-focused abstractive captions dataset is solely to evaluate the quality and alignment of image-text embeddings used for conditional guidance. However, to mitigate any potential risk of these models being used for misinformation generation, we recommend strict guidelines on the responsible use of this technology. This includes using the model only for illustrative purposes and not for creating images that represent real-world events without human oversight. We used LLM-based AI-based proofreading tools solely for minor language and grammar corrections after completing the scientific content. These tools were not used for ideation, writing, analysis, or data generation. All intellectual contributions are those of the authors.

Acknowledgments

Special thanks to Aadarsh Anantha Ramakrishnan for his contributions through insightful research discussions and proofreading of the draft. This research has been partially supported by NSF Awards #1820609 and #2114824.

References

Huggingface sentence-transformers/all-MiniLM-L6-v2. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2. Accessed: 2024-4-5.

Reza Abbasi, Ali Nazari, Aminreza Sefid, Mohammadali Banayeeanzade, Mohammad Hossein Rohban, and Mahdieh Soleymani Baghshah. 2025. Analyzing CLIP's performance limitations in multi-object scenarios: A controlled high-resolution study.

Malihe Alikhani, Sreyasi Nag Chowdhury, Gerard de Melo, and Matthew Stone. 2019. CITE: A corpus of image-text discourse relations. In *Proceedings of the 2019 Conference of the North*, pages 570–575, Stroudsburg, PA, USA. Association for Computational Linguistics.

Malihe Alikhani, Piyush Sharma, Shengjie Li, Radu Soricut, and Matthew Stone. 2020. Cross-modal coherence modeling for caption generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6525–6535, Stroudsburg, PA, USA. Association for Computational Linguistics.

Aashish Anantha Ramakrishnan, Aadarsh Anantha Ramakrishnan, and Dongwon Lee. 2025a. CORDIAL: Can multimodal large language models effectively understand coherence relationships? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21277–21297, Stroudsburg, PA, USA. Association for Computational Linguistics.

Aashish Anantha Ramakrishnan, Aadarsh Anantha Ramakrishnan, and Dongwon Lee. 2025b. IRONIC: Coherence-aware reasoning chains for multi-modal sarcasm detection. In *The First Workshop on Pragmatic Reasoning in Language Models (PragLM)*.

Aashish Anantha Ramakrishnan, Aadarsh Anantha Ramakrishnan, and Dongwon Lee. 2025c. RONA: Pragmatically diverse image captioning with coherence relations. In *Proceedings of the Fourth Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2025)*, pages 74–86. Association for Computational Linguistics.

Amith Ananthram, Elias Stengel-Eskin, Mohit Bansal, and Kathleen McKeown. 2024. See it from my perspective: How language affects cultural bias in image understanding. In *The Thirteenth International Conference on Learning Representations*.

James Betker, Gabriel Goh, Li Jing, and Tim Brooks. Improving image generation with better captions.

Emanuele Bugliarello and Desmond Elliott. 2021. The role of syntactic planning in compositional image captioning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 593–607, Stroudsburg, PA, USA. Association for Computational Linguistics.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3558–3568. IEEE.

Junsong Chen, Y U Jincheng, G E Chongjian, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. 2023. PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *The Twelfth International Conference on Learning Representations*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

N Cohn. 2003. *Visual Syntactic Structures: Towards a Generative Grammar of Visual Language*.

Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. 2022. VQGAN-CLIP: Open domain image generation and editing with natural language guidance. *arXiv:2204.08583 [cs]*.

Damian. compel: A prompting enhancement library for transformers-type text embedding systems.

Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. 2020. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5203–5212. IEEE.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. ArcFace: Additive angular margin loss for deep face recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. 2021. CogView: mastering text-to-image generation via transformers. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA. Curran Associates Inc.

Patrick Esser, Sumith Kulal, A Blattmann, Rahim Entezari, Jonas Muller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. 2024. Scaling rectified flow transformers for high-resolution image synthesis. *ICML*, abs/2403.03206:12606–12633.

Weixi Feng, Wanrong Zhu, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. 2023. LayoutGPT: compositional visual planning and generation with large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Pan Gao, Ying Hu, and Chenyi Zhuang. 2024. Magnet: We never know how text-to-image diffusion models work, until we learn how vision-language models function. In *Advances in Neural Information Processing Systems 37*, volume 37, pages 57115–57149, San Diego, California, USA. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. Mixtral of experts.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. 2022. The role of ImageNet classes in fréchet inception distance.

Daiqing Li, Huan Ling, Seung Wook Kim, Karsten Kreis, Sanja Fidler, and Antonio Torralba. 2022. BigDatasetGAN: Synthesizing ImageNet with pixel-wise annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21330–21340.

Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi, Chaoyou Fu, Ying Tai, Wankou Yang, Yabiao Wang, and Chengjie Wang. 2024a. A survey on benchmarks of multimodal large language models.

Yifan Li, Anh Dao, Wentao Bao, Zhen Tan, Tianlong Chen, Huan Liu, and Yu Kong. 2024b. Facial affective behavior analysis with instruction tuning. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29 – October 4, 2024, Proceedings, Part XVIII*, page 165–186, Berlin, Heidelberg. Springer-Verlag.

Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. 2024. LLM-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *Transactions on Machine Learning Research*.

Jiayi Liao, Xu Chen, Qiang Fu, Lun Du, Xiangnan He, Xiang Wang, Shi Han, and Dongmei Zhang. 2024. Text-to-image generation for abstract concepts. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24, page 9. AAAI Press.

Shanchuan Lin and Xiao Yang. 2023. Diffusion model with perceptual loss.

Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. 2021. Visual news: Benchmark and challenges in news image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6761–6771, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Subhabrata (subho) Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of GPT-4. *arXiv: Computation and Language*.

Son Thanh Nguyen. PyPi package "better_profanity" version 0.7.0.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv:2112.10741 [cs]*.

Mayu Otani, Riku Togashi, Yu Sawai, Ryosuke Ishigami, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Shin'ichi Satoh. 2023. Toward verifiable and reproducible human evaluation for text-to-image generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14277–14286. IEEE.

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. 2021. Learning transferable visual models from natural language supervision. *arXiv:2103.00020 [cs]*.

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

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR.

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

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

Christian Schlarmann, Naman D Singh, Francesco Croce, and Matthias Hein. 2024. Robust CLIP: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. *ArXiv*, abs/2402.12336.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823. IEEE.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, and Others. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294.

Christoph Schuhmann, Robert Kaczmarczyk, Aran Komatsuzaki, Aarush Katta, Richard Vencu, Romain Beaumont, Jenia Jitsev, Theo Coombes, and Clayton Mullis. 2021. LAION-400M: Open dataset of CLIP-filtered 400 million image-text pairs. In *NeurIPS Workshop Datacentric AI*.

Schuhmann C, Köpf A, Vencu R, Coombes T, and Beaumont R. 2022. Laion coco: 600m synthetic captions from laion2b-en. https://laion.ai/blog/laion-coco/.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-OKVQA: A benchmark for visual question answering using world knowledge. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, page 146–162, Berlin, Heidelberg. Springer-Verlag.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2256–2265. PMLR.

Zeliang Song and Xiaofei Zhou. 2021. Exploring explicit and implicit visual relationships for image captioning. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248.

Michael Toker, Hadas Orgad, Mor Ventura, Dana Arad, and Yonatan Belinkov. 2024. Diffusion lens: Interpreting text encoders in text-to-image pipelines. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9713–9728.

Alasdair Tran, Alexander Mathews, and Lexing Xie. 2020. Transform and tell: Entity-aware news image captioning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13032–13042, Seattle, WA, USA. IEEE.

Johannes M van Hulst, Faegheh Hasibi, Koen Dercksen, Krisztian Balog, and Arjen P de Vries. 2020. REL: An entity linker standing on the shoulders of giants. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, pages 2197–2200, New York, NY, USA. Association for Computing Machinery.

Ashish Vaswani, Noam M Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Neural Inf Process Syst*, 30:5998–6008.

Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. 2017. Context-aware captions from context-agnostic supervision. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1070–1079. IEEE.

Jianyi Wang, Kelvin C K Chan, and Chen Change Loy. 2023a. Exploring CLIP for assessing the look and feel of images. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2):2555–2563.

Jiaqi Wang, Zhengliang Liu, Lin Zhao, Zihao Wu, Chong Ma, Sigang Yu, Haixing Dai, Qiushi Yang, Yiheng Liu, Songyao Zhang, Enze Shi, Yi Pan, Tuo Zhang, Dajiang Zhu, Xiang Li, Xi Jiang, Bao Ge, Yixuan Yuan, Dinggang Shen, and 2 others. 2023b. Review of large vision models and visual prompt engineering.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024a. Measuring multimodal mathematical reasoning with MATH-vision dataset. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Qizhou Wang, Yong Lin, Yongqiang Chen, Ludwig Schmidt, Bo Han, and Tong Zhang. 2024b. A sober look at the robustness of CLIPs to spurious features. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yuyang Wang, Anurag Ranjan, Joshua M Susskind, and Miguel Angel Bautista. 2025. INRFlow: Flow matching for INRs in ambient space. In *Forty-second International Conference on Machine Learning*.

Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. 2023c. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 893–911.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. 2023. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis.

Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. 2025. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *International Journal of Computer Vision*, 133(3):1175–1194.

Chunpu Xu, Hanzhuo Tan, Jing Li, and Piji Li. 2022. Understanding social media cross-modality discourse in linguistic space. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2459–2471, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023. ImageReward: learning and evaluating human preferences for text-to-image generation. In

Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1316–1324, Salt Lake City, UT, USA. IEEE.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. 2022. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*.

Ge Yuan, Xiaodong Cun, Yong Zhang, Maomao Li, Chenyang Qi, Xintao Wang, Ying Shan, and Huicheng Zheng. 2023. Inserting anybody in diffusion models via celeb basis.

Yuchen Zeng, Wonjun Kang, Yicong Chen, Hyung Il Koo, and Kangwook Lee. 2024. Can MLLMs perform text-to-image in-context learning? In *First Conference on Language Modeling*.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. Understanding deep learning requires rethinking generalization. *arXiv:1611.03530 [cs]*.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, Salt Lake City, UT. IEEE.

Yue Zhang, Zhiyang Xu, Ying Shen, Parisa Kordjamshidi, and Lifu Huang. 2025. SPARTUN3D: Situated spatial understanding of 3D world in large language model. In *The Thirteenth International Conference on Learning Representations*.

Mingyang Zhou, Grace Luo, Anna Rohrbach, and Zhou Yu. 2022a. Focus! relevant and sufficient context selection for news image captioning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6078–6088.

Victor Zhou. Building a better profanity detection library with scikit-learn. https://victorzhou.com/blog/better-profanity-detection-with-scikit-learn/. Accessed: 2024-9-11.

Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. 2022b. Towards language-free training for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17907–17917. openaccess.thecvf.com.

Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. 2019. DM-GAN: Dynamic memory generative adversarial networks for text-to-image synthesis. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5795–5803, Long Beach, CA, USA. IEEE.

Yongshuo Zong, Ondrej Bohdal, and Timothy Hospedales. 2024. VL-ICL bench: The devil in the details of multimodal in-context learning. In *The Thirteenth International Conference on Learning Representations*.

Appendix

A Dataset Pre-processing

A.1 Image-based Filtering

We standardize the resolution of all images to 512x512. By using Entropy-based cropping, we retain focus on points of interest within a frame. This helps keep the foreground object at the center of the image, limiting information loss to only the background elements. To remove noisy and blurry images, we use CLIP-IQA (Wang et al., 2023a) as a reference-free metric. To filter images based on noisiness and sharpness, we use a minimum threshold of 0.3.

A.2 Caption Filtering and Entity Tagging

In the first stage of filtering, we remove very short captions. We select captions with a minimum length of 6 words and above for our dataset. This is done to ensure that selected captions are informative enough for T2I synthesis. We use different approaches for tagging NEs based on the dataset the captions were extracted from. For captions extracted from the NYTimes800K news corpus, we use the provided NER annotations for filtering. The VisualNews corpus does not provide ground-truth annotations, so we identify mentions of NE using the Spacy library. We remove samples containing 'PERSON', 'GPE', 'LOC', 'WORK_OF_ART', 'ORG' entity types due to their high presence in captions.

B Dataset Insights

B.1 Caption Statistics

In this section, we provide additional statistics on the ANCHOR dataset and analyze the distribution of image-caption pairs. In Table 7, we provide caption statistics of ANCHOR compared to 2 popular image-caption pair datasets: COCO Captions (Chen et al., 2015) and Conceptual Captions 3M (CC3M) (Sharma et al., 2018). By tokenizing and lemmatizing each caption without case sensitivity, we compute the number of unique tokens present in a dataset. We utilize the NLTK library for both tokenization and lemmatization. We can observe that across different data splits of ANCHOR, the mean caption length is significantly higher with a greater variation in caption length compared to other datasets. In addition to the increased caption length, it also contains a significant amount of unique tokens considering the number of samples present. The CLIPScores of samples in ANCHOR are also lower compared COCO Captions, indicating that default CLIP embeddings may struggle in capturing contextual alignment in abstractive captions. This highlights the diversity of captions in ANCHOR, showing greater expression in describing visual concepts.

Algorithm 1: Subject Reweighting Process

Input: Abstractive Caption S_{abstr}, Large Language Model LLM, Tokenizer $\mathcal{T}$, Text Encoder $TextEnc$, T2I Model $\mathcal{G}$, Scale Multiplier β

Output: Generated Image I

Step 1: Tokenization

$\{T_1, T_2, \ldots, T_m\} \leftarrow \mathcal{T}(S_{abstr})$; `// Tokenize caption into subject tokens` T_i

Step 2: Identify Subject Spans

$T_{key} \leftarrow \text{LLM}(\{T_1, \ldots, T_m\})$; `// Pass tokens to the LLM to pick key subjects`

Step 3: Assign Weights

Initialize weight vector W_{abstr} of size m with 1;

for $i = 1$ **to** m **do**

  if $T_i \in T_{key}$ **then**

    $W_{abstr}[i] \leftarrow \beta$; `// Assign upweight` $(\alpha_i = \beta)$ `to key tokens`

Step 4: Generate Original Embeddings

$E_{orig} \leftarrow TextEnc(S_{abstr})$; `// Generate the original/unmodified embeddings`

Step 5: Apply Weights

$E_{abstr} \leftarrow E_{orig} \odot W_{abstr}$; `// Multiply each token embedding by its weight`

Step 6: Image Generation

$I \leftarrow \mathcal{G}(E_{abstr})$; `// Use reweighted embeddings as guidance for T2I`

return I;

B.2 Article Topic Distribution

For analyzing the categories of articles from which image-caption pairs have been selected for our dataset, we provide a unified category list in Figure 5. With articles sourced from different news agencies, each source has its own article category

Dataset	Unique Tokens	CLIPScore (↑)	Caption Length	
			Mean	StdDev
COCO Captions Train	22767	0.5152	10.42	0.88
COCO Captions Val	16647	0.5237	10.42	0.87
CC3M Train	45896	0.3984	10.31	3.30
CC3M Val	9289	0.4880	10.40	3.35
ANCHOR Non-Entity Train	51026	0.4797	14.84	5.51
ANCHOR Non-Entity Val	10619	0.4811	14.93	5.38
ANCHOR Non-Entity Test	10485	0.4807	14.67	5.36
ANCHOR Entity	10955	0.4960	**22.13**	**7.95**

Table 7: Caption Statistics of ANCHOR. Overall, we observe the captions from ANCHOR are more lexically and semantically diverse compared to traditional T2I evaluation benchmarks.

taxonomy. To create a unified taxonomy, we fix the categories provided by articles from NYTimes as our template. To cluster similar article categories under one label, we utilize the lightweight sentence transformer all-MiniLM-L6-v2 (Unk). With a minimum similarity threshold of 0.5, we cluster every sample's default topic description into NYTimes's taxonomy labels. Here, we visualize our dataset's top 30 article classes, showing the diverse spread of image-caption pairs present.

GPT-3.5/4 & Mixtral 8x7B Prompt

User: Use only the information provided in the prompt for answering the question. List the main topic word and additional topic words from the given image caption in the format: {"main_topic_word": <insert-topic-word-string>, "additional_topic_words": [<insert-topic-word1>, ...]}. Caption Text: <insert-caption-text>

Orca Mini 13B Prompt

User: "User: List only the main objects from the sentence: <insert-caption-text>"

Caption-Rewriting Prompt

User: Write a simple prompt for an image generation model to generate an image for the given text: <insert-caption>

C Salient Subject Selection

The detailed overview on how we perform Subject weighting using the Compel library (Damian) is described in Algorithm 1. To reduce memory requirements and to speed up inference, we initialize in mixed precision mode and set $dtype = float16$. The system prompt is set as "*You are an AI assistant that follows instructions extremely well. Help as much as you can.*" In cases where the LLM returned no salient subject phrases, we use only the text caption without any subject weights. Given the size of our dataset (70K+ image-caption pairs), GPT 3.5 Turbo offered the best balance of consistency, latency, and token cost, making it our preferred backbone to run the majority of our ablation studies.

D Implementation Details

D.1 Metric Selection and Justification

Frechet Inception Distance (Heusel et al., 2017) serves as an indicator to quantify the overall realism and diversity of generated samples compared to the ground truth images. With the distribution of datasets like ANCHOR diverging significantly from the Inception-V3 used in traditional FID calculations (Kynkäänniemi et al., 2022), we adopt the more representative FID_{CLIP} metric for our testing. To measure the relatedness of our generated images and ground truth captions, we utilize IR. Compared to image-caption similarity metrics like CLIPScore (Hessel et al., 2021), IR is trained on real-world image-caption pairs annotated and ranked according to human preference. Similarly, Human Preference Score V2 also serves as an indicator of human preference alignment. For Face Detection, we utilize a RetinaFace-based detector and measure the average number of times a face is detected across generated images. The ArcFace

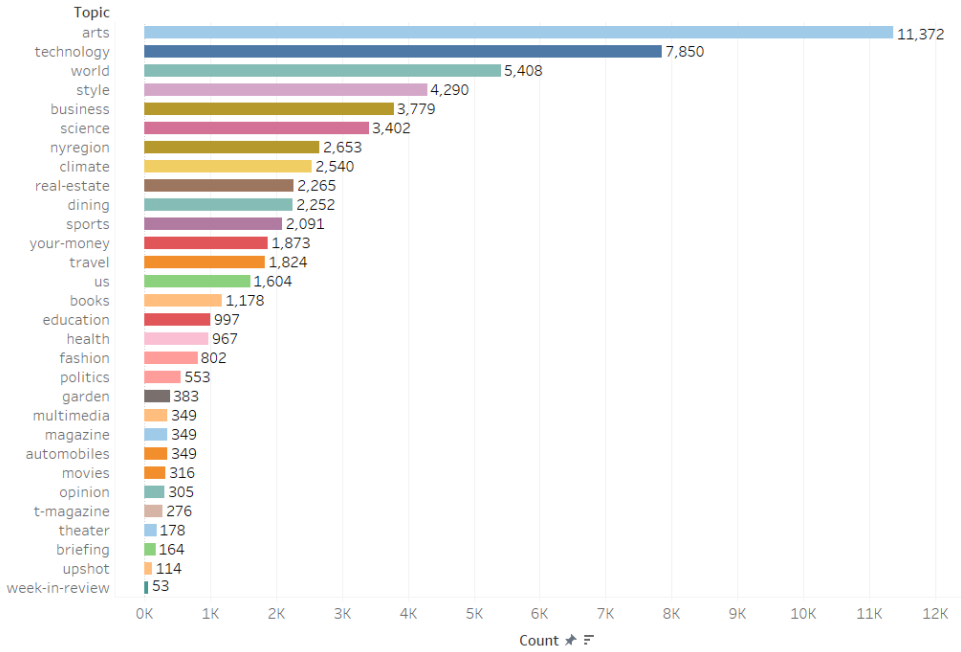


Figure 5: Distribution of Article Topics for samples in ANCHOR.

(Deng et al., 2019) face recognition model is used for calculating Identity Preservation scores.

D.2 Model Training and Inference

We set *guidance_scale = 7.5* and *num_inference_steps = 100*. All our reported metrics are averaged across 2 random seeds 42 and 3. All generated images are in 512x512 resolution. Our SAFE Model has been fine-tuned for 300 epochs. We utilize the same inference hyper-parameters as the baseline Stable Diffusion models. A learning rate of $5 * 10^{-5}$ was set for SAFE. All experiments were performed on a Nvidia A100 GPU taking up to 200 GPU hours to run. To perform prompt weighting, after experimenting with different weight scales, we apply a uniform increase of x2 or $(1.1)^2$ for all LLM extracted keywords in the original caption. Generated samples of Baseline and SAFE models using the same seed are presented in Figure 4. We select GPT-3.5 as our default LLM model for collecting subject weights for all our fine-tuned models. On the ANCHOR Entity test-set, we assess the impact subject conditioning has in understanding abstractive captions containing NEs.

E Qualitative Evaluation of Generated Samples

For this project, all human evaluation surveys were created on Qualtrics and distributed through Amazon MTurk with our survey UIs provided in Figures 7, 8. All our studies have been conducted with Institutional Review Board (IRB) approval. We do not collect any personally identifiable data from participants in our study. Voluntary consent is obtained from each participant before taking part in all studies. We provide clear instructions for each evaluation task presented to participants with examples and test their understanding using a pre-survey questionnaire. This is done to ensure data quality and improve the consistency of task understanding across participants. Attention Check questions were also incorporated to prevent low-quality submissions from being accepted. The demographic for participants taking part in our survey was limited to people above 18 years of age.

E.1 Qualitative Evaluation Setup for SAFE

We perform human evaluation of SAFE vs baseline Stable Diffusion on Amazon MTurk to understand perceived variations in subject understanding. From the ANCHOR Non-Entity test set, we randomly sample and filter 300 captions for our survey. The questions in our survey require participants to pick the image that is most related to the provided caption and also rate the difficulty of choosing between the two images on a 5-point scale. The scale ranges from 1 - "Very easy to distinguish" to 5 - "Very difficult to distinguish". This measure is utilized to understand the rater's

Model	Identity (↑)	Detect (↑)	FID_{CLIP} (↓)	IR (↑)	HPS V2 (↑)
SAFE (DFE + GPT-3.5)	0.3323	0.9498	30.7756	**0.6072**	0.2564
Stable Diffusion 2.1 (Base)	**0.3391**	**0.9533**	**30.0651**	0.6060	**0.2565**

Table 8: Results of Abstractive Text-to-Image synthesis on ANCHOR Entity Test-set averaged across all classes. As expected, the introduction of ANCHOR results in a drop in generation quality of NE features while retaining image-caption alignment. This supports our hypothesis that Non-Entity and Entity features are represented differently by encoders such as CLIP.

Figure 6: Qualitative comparison of different T2I models on ANCHOR Entity Set. Words highlighted in Orange are used for subject conditioning. Although images generated using SAFE improve on multi-subject representations, the accurate generation of visual features of NEs still serves as a challenge.

confidence in assessing the image-caption pairs. We removed all samples rated as "Very difficult to distinguish" from our analysis to ensure the rating confidence. With each rater labeling a max of 10 samples, we removed all submissions where raters failed our attention checks. Thus, a min of 3 and

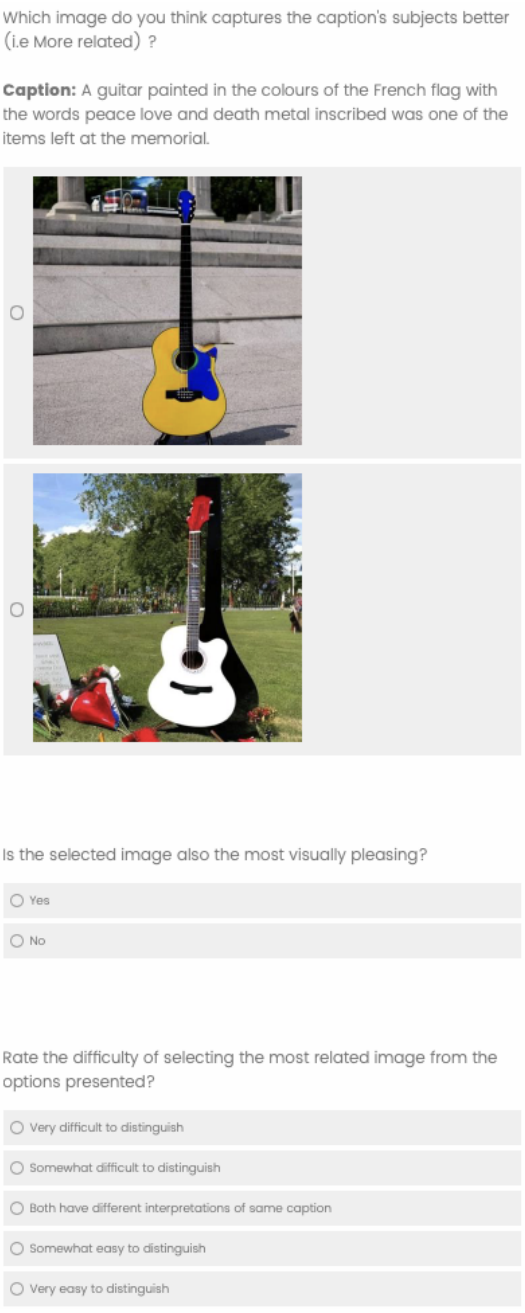


Figure 7: Survey UI for Generated Image Evaluation Study

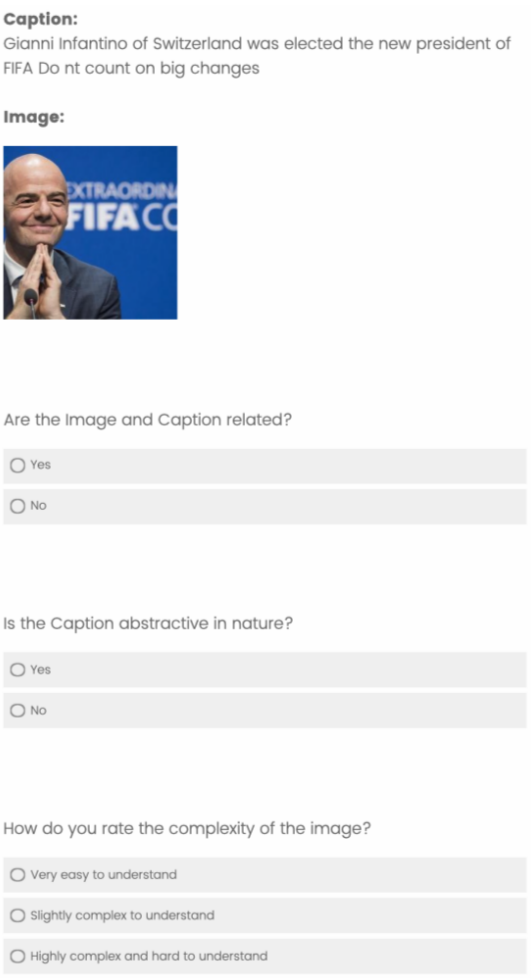


Figure 8: Survey UI for Data Quality Evaluation Study

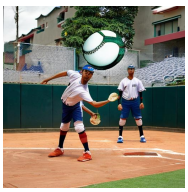
a max of 5 annotations per sample were present in our final evaluation set. Our analysis shows that raters consistently preferred images generated by SAFE over the baseline model, complementing our quantitative results.

E.2 Inter-Annotator Agreement

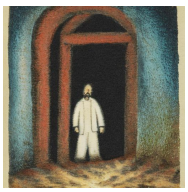
Additionally, we compute the inter-annotator agreement scores for our MTurk participants to assess the significance of our results. For our human evaluation experiments with SAFE, we use the Krippendorff's α metric which shows an inter-annotator agreement of $\alpha = 0.1216$. This shows a positive correlation between annotators concerning the collected ratings for this task. With the absolute score of α being lower than the average scores reported on other rating tasks, we identify key reasons why this may be the case. In our case, each annotator does not rate every question present in our evaluation samples. So, the unanswered questions by a survey participant are treated as missing values. The high number of missing values when utilizing the typical formulation of this metric is one reason for the lower score observed. Additionally, other studies attempting to assess inter-annotator agreement of T2I generators (Otani et al., 2023) on complex text-image datasets such as DrawBench (Saharia et al., 2022) have reported similarly low scores, indicating the difficulty of this task.

F Additional Examples

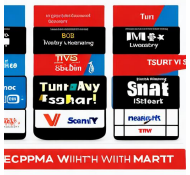
We provide additional generated examples using both baseline and SAFE models for reference. Figure 9 with examples Ex10, Ex11, Ex12 are generated from the test set of ANCHOR Non-Entity. Similarly Ex7 and Ex8 from Figure 6 are examples from ANCHOR Entity.

Original	SD 2.1 (Base)	SD 2.1 (CR)	SAFE (DFE + GPT-3.5)
			

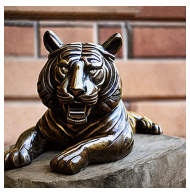
Ex8: At first glance, pelota mixteca might resemble elements of baseball, volleyball and tennis, but a closer examination reveals a bit more nuance. Each jugada, as each individual game is called, involves approximately 10 players, and begins when one player initiates a serve from the cement slab.

Ex9: The painted figure of a man is illuminated through a doorway to the dwelling.

Ex10: Mediastreaming boxes can turn any TV smart or add features and channels to others for as little as 15.

Ex11: A bronze tiger shows assertiveness and a winning spirit. The books are all from colleagues.

Ex12: A sharp knife, one of a cook's essential tools, is used to carefully cut onions, which are easier to brown (if they're not bludgeoned) for a confit.

Figure 9: Qualitative comparison of different T2I models on ANCHOR Non-Entity Subset. Words highlighted in Orange are used for subject conditioning