REFVNLI: Towards Scalable Evaluation of Subject-driven Text-to-image Generation

Aviv Slobodkin[1*] **Hagai Taitelbaum**[1] **Yonatan Bitton**[1]
Brian Gordon[1*] **Michal Sokolik**[1] **Nitzan Bitton Guetta**[2] **Almog Gueta**[1]
Royi Rassin[1*] **Dani Lischinski**[1] **Idan Szpektor**[1]
[1]Google Research [2]Ben Gurion University
{slobodkin,hagait,yonatanbitton}@google.com

Abstract

Subject-driven text-to-image (T2I) generation aims to produce images that align with a given textual description, while preserving the visual identity from a referenced subject image. Despite its broad downstream applicability—ranging from enhanced personalization in image generation to consistent character representation in video rendering—progress in this field is limited by the lack of reliable automatic evaluation. Existing methods either assess only one aspect of the task (i.e., textual alignment or subject preservation), misalign with human judgments, or rely on costly API-based evaluation. To address this gap, we introduce REFVNLI, a cost-effective metric that evaluates both textual alignment and subject preservation in a single run. Trained on a large-scale dataset derived from video-reasoning benchmarks and image perturbations, REFVNLI outperforms or statistically matches existing baselines across multiple benchmarks and subject categories (e.g., *Animal*, *Object*), achieving up to 6.4-point gains in textual alignment and 5.9-point gains in subject preservation.[1]

1 Introduction

In a well-known scene from *"The Little Prince"*, the narrator attempts to comfort a grieving prince by saying *"I'll draw you a fence around your flower"*. While fairly simple, this offer raises a deeper question: what makes such a drawing adequate? Beyond accurately depicting a fence around a flower, the use of *'your'* implies that it must portray a specific flower—the Prince's own— one with which he shares a history. Given the flower's uniqueness and distinct visual traits, the narrator's task proves far more complex than it first appears.

Subject-driven text-to-image (T2I) generation (Chen et al., 2024; Li et al., 2024; Ruiz et al., 2023) enables a variety of downstream applications, such as personalized image generation (Ruiz et al., 2023), character consistency in video generation (Liu et al., 2024c), and enhancing vanilla T2I evaluation frameworks for less-known entities via image retrieval (Tahmasebi et al., 2025). Unlike standard T2I models that are only conditioned on text inputs, this setup takes both a textual prompt and a reference image, enabling more precise subject representation. For example, when creating an image of a fenced-in flower for the Little Prince, subject-driven models should use a reference image of the Prince's flower to ensure the output preserves its unique features (see Fig. 1).

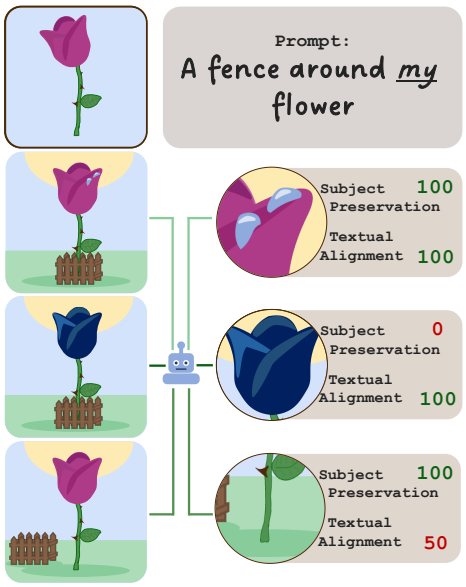

Figure 1: **Illustration of REFVNLI:** Given a reference image of a subject, a *prompt* referring to the subject, and a target image, REFVNLI assesses both subject preservation and textual alignment. For **subject preservation**, it distinguishes identity-preserving variations, like dew on a flower (top image), from identity-altering changes, such as color change (middle image). For **textual alignment**, it assesses whether the target image reflects all details from the *prompt*, such as the fence's position relative to the flower (bottom image).

[*] Work done during an internship at Google Research.
[1] https://google.github.io/refvnli/

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 8420–8438
November 4-9, 2025

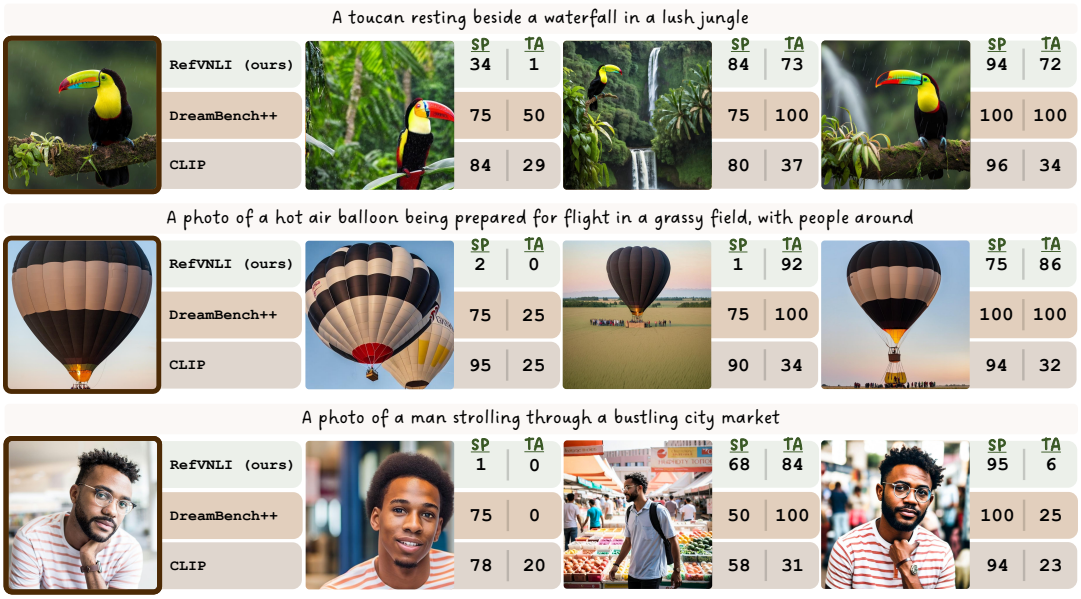

Figure 2: **Qualitative Comparison**: We compare REFVNLI with DreamBench++ and CLIP, which score both **Subject Preservation (SP)** and **Textual Alignment (TA)**, using examples from the *Animal*, *Object*, and *Human* categories. DreamBench++ scores (0-4) are scaled to 0-100 for better readability. REFVNLI exhibits better robustness to identity-agnostic changes (SP), such as the zoomed-out parrot (top-middle) and the zoomed-out person with different attire (bottom-middle). It is also more sensitive to identity-defining traits, penalizing changed facial features (left-most person) and mismatched object patterns (left and middle balloons). Additionally, REFVNLI excels at detecting text-image mismatches (TA), as seen in its penalization of the top-left image for lacking a waterfall.

Despite its wide applicability, research in this area has been hindered by a lack of reliable and scalable auto-raters. Existing metrics typically correlate poorly with humans, and often focus only on *textual alignment* between the input prompt and the target image, as in CLIP-T (Radford et al., 2021) and SigLIP (Zhai et al., 2023), or on *subject preservation* between the input and target images, as in CLIP-I (Radford et al., 2021) and DINO-I (Caron et al., 2021), while both aspects are needed for successful subject-driven generation. More correlative metrics, like DreamBench++ (Peng et al., 2024) and VIEScore (Ku et al., 2024a), depend on costly API calls to models like GPT-4 (OpenAI, 2024), making them less scalable and reproducible.

To bridge this gap, we present REFVNLI, a cost-effective fine-tuned auto-rater for subject-driven T2I generation. Given a triplet $<image_{ref}, prompt, image_{tgt}>$, REFVNLI predicts two scores—*textual alignment* and *subject preservation*—in a single run, as shown in Fig. 1. To train REFVNLI, we automatically curate a large-scale dataset of $<image_{ref}, prompt, image_{tgt}>$ triplets, labeled with $<$*textual alignment, subject preservation*$> \in \{0,1\}^2$. For subject preservation, we identify subjects across video frames, creating positive examples using pairs of frames depicting the same subject, and negative ones by pairing frames of different subjects (yet of the same entity). This approach enables robustness to variations in subject appearance (e.g., rotation, setting, clothing), as well as to the presence of extraneous elements (e.g., dew on the Little Prince's flower in Fig. 1, top). At the same time, REFVNLI must also be sensitive to identity-defining traits, such as human facial features or object shapes and colors (e.g., middle image in Fig. 1). To this end, we modify images by masking and inpainting identity-critical regions, while keeping everything else unchanged. The original subject crops are then paired with the unaltered images as positive pairs and with the modified images as negative pairs, thereby teaching the model to focus on key identity attributes.

For textual alignment, we first create positive *image-prompt* pairs. For that, we use an LLM to caption each image in the aforementioned pairs, ensuring focus on the subject by enclosing it within a bounding box. Negative pairs are then formed by replacing these captions with those of different scenes. For extra sensitivity to minor mismatches, like a fence drawn *next to* rather than *around* the Prince's flower (Fig. 1, bottom), we also create negative pairs by altering a single fact in each original (positive) caption. Finally, to derive the $<image_{ref},$

Figure 3: Generating subject preservation classification training instances from video frames. Given two pairs of frames, each extracted from distinct video scenes featuring the same entity (e.g., *a* dog), where both frames within each pair depict the same subject (e.g., *the same* dog), we construct training {$image_{ref}$, $image_{tgt}$} pairs for subject preservation classification. **Positive pairs** are formed by pairing a cropped subject from one frame (e.g., dog from left frame in Scene 1) with the full frame from the same scene (right frame in Scene 1). In contrast, **negative pairs** are created by pairing the cropped subject with the other scene's full frames (e.g., Scene 2). This process is applied to all four frames, with each taking turns as the cropped reference image ($image_{ref}$), while the corresponding full-frame counterparts serve as $image_{tgt}$, yielding a total of 4 positive and 8 negative training pairs.

prompt, $image_{tgt}$> triplets from each pair of frames and associated captions, we use the cropped subject from one frame as $image_{ref}$ and the entire second frame, alongside its caption, as {*prompt*, $image_{tgt}$}, resulting in a total of 1.2 million instances.

We evaluate REFVNLI on multiple human-labeled test sets for subject-driven generation, including DreamBench++ (Peng et al., 2024), ImagenHub (Ku et al., 2024b), and KITTEN (Huang et al., 2024), across categories such as *Humans*, *Animals*, *Objects*, *Landmarks*, and a multi-subject setting. For **textual alignment**, REFVNLI consistently matches or outperforms all baselines, with up to 6.4-point gains in *Landmarks* and proficiency at detecting subtle text-image misalignments (e.g., missing waterfall in Fig. 2, top-left). It also leads in **subject preservation**, with gains of up to 6.3 points on *Objects* (Table 1) and 5.9 points in the multi-subject setting, surpassing the larger GPT-4o-based DreamBench++ baseline. As seen in Fig. 2, it balances robustness to non-critical changes (e.g., zoomed-out toucan, top-middle) with sensitivity to identity shifts (e.g., altered facial features, bottom-left). Further, REFVNLI effectively handles rare subjects (§5), outperforming all baselines and highlighting its value as a reliable alternative to standard T2I metrics for uncommon entities.

2 REFVNLI: Automatic Metric for Subject-driven T2I Generation

We introduce REFVNLI, a cost-effective auto-rater specifically tailored for subject-driven T2I generation. This section details the automated pipeline used to construct its training dataset (§2.1) and the subsequent training process of REFVNLI (§2.2).

2.1 Training Dataset Construction

To train REFVNLI, we collect a large scale dataset of <$image_{ref}$, *prompt*, $image_{tgt}$> triplets, each with two binary labels: one for **subject preservation** of $image_{ref}$ in $image_{tgt}$, and one for **textual alignment** between the *prompt* and $image_{tgt}$. This involves first creating subject-driven {$image_{ref}$, $image_{tgt}$} pairs, followed by automatic generation of subject-focused *prompts* for each $image_{tgt}$.

Subject-driven image pairs. To ensure our {$image_{ref}$, $image_{tgt}$} dataset is robust to identity-agnostic changes (e.g., pose, clothing, or lighting changes), we use video-based datasets that inherently capture these differences. Specifically, we use Mementos (Wang et al., 2024), comprising scene-specific video frames with human-written textual descriptions, and TVQA+ (Lei et al., 2020), containing human-annotated bounding boxes for characters and objects in TV episodes. We first locate subjects within frames: for Mementos, we extract entities from the provided textual descriptions using Gemini (Team, 2024) and localize them in the associated frames with an object detection model (Minderer et al., 2022), while for TVQA+, we directly use the provided bounding boxes. **Positive pairs** are formed from frames featuring the same subject, usually within the same scene,[2] while **negative pairs** consist of frames with distinct subjects

[2]For TVQA+, we also include cross-scene positive pairs for named entities, such as TV characters.

Figure 4: Creating identity-sensitive {$image_{ref}$, $image_{tgt}$} pairs. Starting with an image and a mask of a subject (e.g., a bag), we randomly keep 5 patches within the masked area ([1]) and use them to create 5 inpainted versions ([2]). The version with the highest MSE between the altered and original areas (e.g., bottom image, MSE = 3983) is paired with the *unmodified* crop to form a **negative pair**, while the original image and the same crop create a **positive pair**, with the crop acting as $image_{ref}$ in both cases.

(of the same type of entity), often across scenes (see Fig. 3). These *frame*-pairs are then converted into {$image_{ref}$, $image_{tgt}$}-pairs by cropping the subject from one frame as $image_{ref}$ (e.g., left frame in Fig. 3, Scene 1) and using the full second frame as $image_{tgt}$ (each of the other frames in Fig. 3). This is then repeated with reversed roles for an extra {$image_{ref}$, $image_{tgt}$} pair. In total, we collected 338,551 image pairs (228,661 from Mementos and 109,890 from TVQA+) from 44,418 unique frames.

To further enhance sensitivity to identity-specific attributes, such as facial features in humans or shapes and patterns in objects, we leverage the Open Images dataset (Kuznetsova et al., 2020) to create additional training instances, as shown in Fig. 4. Using its gold segmentation masks, we selectively mask and inpaint identity-critical regions while preserving other details. Specifically, we randomly sample 5 sub-masks covering 30%-50% of the subject mask ([1] in Fig. 4), which we use to create 5 inpainted variants ([2]). The version with the highest Mean Squared Error (MSE) between the modified and original regions (e.g., Fig. 4, bottom image, MSE=3983) is then paired with the *unmodified* cropped subject to form a **negative pair** of {$image_{ref}$, $image_{tgt}$}, while the original image and the same crop form a **positive pair**, with the crop serving as $image_{ref}$ in both cases. This process yields extra 16,572 pairs, helping the model focus on fine-grained identity details. To further improve data quality, we also apply multiple filtering steps, including removing blurry images and those with unclear subjects (see Appendix C.2 for details).

Image-*prompt* pairs. For each {$image_{ref}$, $image_{tgt}$} pair, we generate positive and negative *prompts* for $image_{tgt}$ (Fig. 5). **Positive *prompts*** (Fig. 5, top) are created by instructing Gemini (Team, 2024) to describe $image_{tgt}$, ensuring the subject is explicitly mentioned by enclosing it in a bounding box and guiding the model to focus on it, as well as filtering out *prompts* lacking it. For **negative *prompts*** (Fig. 5, middle), we swap *prompts* between frames containing the same entity type (e.g., a dog). To further enhance sensitivity to subtle mismatches, we also create **hard-negative *prompts*** (Fig. 5, bottom) by using Gemini to modify a single non-subject detail in the positive *prompts*, following Gordon et al. (2024). In total, this and the image-pairing steps yield 1.2 million <$image_{ref}$, *prompt*, $image_{tgt}$> triplets labeled for *textual alignment* and *subject preservation*.

2.2 REFVNLI Training

We fine-tune PaliGemma (Beyer et al., 2024), a 3B Vision-Language Model (VLM) known for effective transfer learning, focusing on a variant adapted for multi-image inputs.[3] The model takes as input two images ($image_{ref}$ and $image_{tgt}$), and a *prompt*

[3] https://huggingface.co/google/paligemma-3b-ft-nlvr2-448

Figure 5: Example of *prompt-image*$_{tgt}$ pairs. Given an image with some subject (e.g., a dog), we create a **positive *prompt*** by adding a bounding box around the subject and directing Gemini to describe it (top *prompt*s). **Negative *prompt*s** are created by swapping *prompt*s between images of the same entity (middle *prompt*s). For additional **hard negatives**, we guide Gemini to modify a single non-subject detail in the positive *prompt* while keeping the rest unchanged (bottom *prompt*s).

that includes <*u*> and <*u*> markups around the referenced subject. During training, the model performs two sequential binary classifications—first assessing textual alignment, then subject preservation—outputting *'1'* (positive) or *'0'* (negative) for each task. At inference, we compute the probabilities of predicting *'1'* and *'0'* for the first and second generated tokens, and use their ratio to calculate the textual alignment and subject preservation scores, respectively. [4]

3 Experimental Settings

This section outlines our meta-evaluation protocol and benchmarks (§3.1), followed by an overview of the baseline models used for comparison (§3.2).

3.1 Meta-evaluation and Benchmarks

We include 3 subject-driven generation benchmarks with human annotations for textual alignment and subject preservation across categories such as *Human*, *Animal*, *Object*, and *Landmark*. To enable a unified evaluation framework, given differing scoring methods (5-scale and binary), we convert all annotations into binary labels: one for whether *image*$_{tgt}$ fully captures the *prompt* (textual alignment) and another for whether it correctly depicts the referenced subject (subject preservation).

For meta-evaluation, we report ROC AUC for each criterion, following standard practice (Honovich et al., 2022; Yarom et al., 2023; Zha et al., 2023), and also compute a unified score as the harmonic mean of the two scores. Following Honovich et al. (2022), significance testing is assessed via bootstrap resampling (Efron, 1987), comparing each baseline to REFVNLI. We report mean scores and highlight models with statistically significant under- or outperformance relative to REFVNLI.

We next present the 3 analyzed benchmarks.

Dreambench++ (Peng et al., 2024) is a subject-driven generation benchmark with human annotations for 8,190 images generated by 7 models. Annotators rated textual alignment and subject preservation on a 0-4 scale, with each image evaluated by 2 raters. To convert these ratings into binary labels, we classify a criterion as positive if both scores are at least 3, and at least one is a 4. We report performance separately for the benchmark's three subject categories: **Human**, **Animal**, and **Object**.[5]

ImagenHub (Ku et al., 2024b) is a human-annotated benchmark for conditional image generation, covering a subject-driven task (150 instances) and a multi-concept task (102 instances), which involves 2 referenced subjects per instance. Each image was rated by 3 annotators. Instead of separate ratings for textual alignment and subject preservation, annotators provided a single adherence score (0, 0.5, or 1) per image. To align with our binary labeling framework, images rated 1 by all 3 annotators were assigned positive labels for both criteria, while the rest were re-annotated by this paper's authors. In the Multi-subject setting, a positive subject preservation label was assigned only when both subjects were accurately depicted. For evaluation, we report separate scores for **Animals** and **Objects** in the single-subject task, while **Multi-subject** instances are split into two single-subject evaluations, with the final score being the lower rating (per criterion), to ensure a stricter assessment.

KITTEN (Huang et al., 2024) evaluates subject-driven T2I models on generating diverse real-world entities (e.g., plants, vehicles, landmarks), using 5 reference images and a *prompt*. Annotators rated entity depiction on a 1–5 scale and provided binary textual alignment scores, with each image assessed by 5 annotators. Unlike our focus on *specific subjects*, KITTEN evaluates *general entity alignment* (e.g., a generic rose rather than a specific one).

[4] See Appendices A and D for more details and ablations.

[5] A fourth 'style' category is excluded as it is beyond our work's scope.

	Textual Alignment			Subject Preservation			Unified Evaluation		
	Animal	Human	Object	Animal	Human	Object	Animal	Human	Object
CLIP	72.8$^{\downarrow}$	77.4$^{\downarrow}$	74.6$^{\downarrow}$	72.4$^{\downarrow}$	**87.7**	76.4$^{\downarrow}$	72.6$^{\downarrow}$	82.2	75.5$^{\downarrow}$
DINO	-	-	-	**80.1**	78.0$^{\downarrow}$	77.3$^{\downarrow}$	-	-	-
Crop-IR	-	-	-	76.9$^{\downarrow}$	85.6	83.4	-	-	-
ArcFace	-	-	-	-	61.0$^{\downarrow}$	-	-	-	-
CLIPScore	71.5$^{\downarrow}$	76.1$^{\downarrow}$	72.9$^{\downarrow}$	-	-	-	-	-	-
BLIPScore	75.4$^{\downarrow}$	79.5$^{\downarrow}$	78.9$^{\downarrow}$	-	-	-	-	-	-
SigLIP	72.5$^{\downarrow}$	80.2	77.1$^{\downarrow}$	-	-	-	-	-	-
TIFA	70.6$^{\downarrow}$	75.7$^{\downarrow}$	69.5$^{\downarrow}$	-	-	-	-	-	-
VQAScore	79.4	78.0$^{\downarrow}$	**82.6**	-	-	-	-	-	-
VIEScore	77.9	75.2$^{\downarrow}$	73.3$^{\downarrow}$	63.4$^{\downarrow}$	81.1	76.4$^{\downarrow}$	69.8$^{\downarrow}$	77.7	74.8$^{\downarrow}$
DreamBench++	79.5	**82.7**	82.5	74.5$^{\downarrow}$	84.1	79.4$^{\downarrow}$	76.9$^{\downarrow}$	83.4	80.9$^{\downarrow}$
PaliGemma$_{\text{text/ref}}$	77.9$^{\downarrow}$	79.2$^{\downarrow}$	81.2	70.1$^{\downarrow}$	71.2$^{\downarrow}$	77.6$^{\downarrow}$	73.8$^{\downarrow}$	74.9$^{\downarrow}$	79.4$^{\downarrow}$
RefVNLI	**80.2**	82.5	82.0	79.4	86.0	**85.7**	**79.8**	**84.2**	**83.8**

Table 1: **ROC AUC scores on DreamBench++** for textual alignment, subject preservation, and their harmonic mean (as a unified evaluation) across *Animal*, *Human*, and *Object* categories. The last two rows feature models finetuned on our dataset, with PaliGemma$_{\text{text/ref}}$ comprising two separate models (PaliGemma$_{\text{text}}$ and PaliGemma$_{\text{ref}}$) trained exclusively for each criterion. Bold indicates the highest score per column. $^{\downarrow}$ and $^{\uparrow}$ indicate statistically significant underperformance and outperformance relative to RefVNLI, respectively.

Hence, we only use the 256 **Landmark** images, as landmarks are unique entities where *entity* adherence coincides with *subject* adherence. To convert ratings into binary labels, we apply majority voting for textual alignment and consider subject preservation positive only if most annotators rated it at least 4 and the average score is 4 or higher.

3.2 Baselines

We evaluate RefVNLI against both standard and state-of-the-art methods for measuring textual alignment, subject preservation, or both.

Baselines for textual alignment. We compare RefVNLI with two groups of automatic metrics for textual alignment. The first group leverages large vision-language models (VLMs), computing cosine similarity between text and image encodings. This includes BLIPScore (Li et al., 2022), CLIPScore (Hessel et al., 2021), and SigLIP (Zhai et al., 2023). The second group, which includes TIFA (Hu et al., 2023) and VQAScore (Lin et al., 2024), evaluates textual alignment via visual question answering (VQA). We also include a baseline where PaliGemma is finetuned on our dataset exclusively for textual alignment, given only the prompt and target image, referred to as PaliGemma$_{\text{text}}$.

Baselines for subject preservation. For subject preservation, we compare RefVNLI to baselines that use large VLMs by computing cosine similarity between reference and target image embeddings. These include DINO (Caron et al., 2021), Crop-IR (Winter et al., 2024),[6] and for the *Human* category, also ArcFace (Deng et al., 2019), a face-recognition model. We also assess a PaliGemma model finetuned on our dataset solely for subject preservation, using only reference and target images (formatted as in §2.2), denoted as PaliGemma$_{\text{ref}}$.

Baselines for both criteria. We also include 3 metrics that assess both criteria. CLIP (Radford et al., 2021) computes scores separately for each criterion by calculating cosine similarity between the encodings of $image_{tgt}$ and those of *prompt* and $image_{ref}$. VIEScore (Ku et al., 2024a) uses an elaborate GPT-4o (OpenAI, 2024) few-shot strategy, simultaneously generating two 0–10 ratings, one for each criterion. Lastly, DreamBench++ (Peng et al., 2024) evaluates each criterion separately using distinct GPT-4o prompts with hand-crafted instructions and examples. This method follows a two-step prompting process, where GPT-4o first summarizes the evaluation task to increase task comprehension before assigning a 0–4 score.

4 Results

Our main results are summarized in Tables 1, 2, and 3, with qualitative examples in Fig. 2.

On DreamBench++ (Table 1), RefVNLI outperforms or statistically matches all baselines across both criteria, with a notable 6.3-point lead over the GPT-4o-based DreamBench++ metric in *subject preservation* for *Objects*. This is especially notable given the benchmark's diverse visual styles,

[6] See Appendix E for more details.

	Textual Alignment			Subject Preservation			Unified Evaluation		
	Animal	Object	Multi-subj.	Animal	Object	Multi-subj.	Animal	Object	Multi-subj.
CLIP	81.8	74.7$^{\downarrow}$	81.1$^{\downarrow}$	63.8$^{\downarrow}$	73.3$^{\downarrow}$	52.6$^{\downarrow}$	71.6$^{\downarrow}$	74.0$^{\downarrow}$	63.8$^{\downarrow}$
DINO	-	-	-	81.7	77.3$^{\downarrow}$	50.0$^{\downarrow}$	-	-	-
Crop-IR	-	-	-	77.6	**84.1**	56.8	-	-	-
CLIPScore	81.5	75.0$^{\downarrow}$	79.1$^{\downarrow}$	-	-	-	-	-	-
BLIPScore	82.9	79.7$^{\downarrow}$	84.2$^{\downarrow}$	-	-	-	-	-	-
SigLIP	80.7	80.6$^{\downarrow}$	82.3$^{\downarrow}$	-	-	-	-	-	-
TIFA	79.9	76.1$^{\downarrow}$	79.2$^{\downarrow}$	-	-	-	-	-	-
VQAScore	77.3$^{\downarrow}$	83.8$^{\downarrow}$	87.8	-	-	-	-	-	-
VIEScore	62.1$^{\downarrow}$	54.1$^{\downarrow}$	71.6$^{\downarrow}$	56.4$^{\downarrow}$	49.4$^{\downarrow}$	50.2$^{\downarrow}$	59.0$^{\downarrow}$	51.5$^{\downarrow}$	58.9$^{\downarrow}$
DreamBench++	**86.4**	85.5$^{\downarrow}$	**88.2**	71.1$^{\downarrow}$	84.0	54.3$^{\downarrow}$	78.0$^{\downarrow}$	84.8	67.2$^{\downarrow}$
PaliGemma$_{\text{text/ref}}$	81.1	88.1	85.3	**82.0**	74.2$^{\downarrow}$	62.1	81.5	80.5$^{\downarrow}$	71.8
REFVNLI	84.6	**89.4**	86.2	80.2	83.8	**62.7**	**82.3**	**86.5**	**72.6**

Table 2: **ROC AUC scores on ImagenHub** for textual alignment, subject preservation, and their harmonic mean (as a unified evaluation) across *Animal* and *Object* categories, as well as for the *Multi-subject* setting.

	Textual Alignment	Subject Preservation	Unified Evaluation
CLIP	83.2$^{\downarrow}$	80.1	81.5$^{\downarrow}$
DINO	-	85.4	-
Crop-IR	-	**90.2**$^{\uparrow}$	-
CLIPScore	83.3$^{\downarrow}$	-	-
BLIPScore	82.6$^{\downarrow}$	-	-
SigLIP	75.3$^{\downarrow}$	-	-
TIFA	90.6$^{\downarrow}$	-	-
VQAScore	89.0$^{\downarrow}$	-	-
VIEScore	82.5$^{\downarrow}$	87.5	84.9
DreamBench++	87.0$^{\downarrow}$	89.9$^{\uparrow}$	88.4
PaliGemma$_{\text{text/ref}}$	94.5	87.5	**90.8**
REFVNLI	**97.0**	82.2	88.9

Table 3: **ROC AUC scores on KITTEN (landmarks)** for textual alignment, subject preservation, and their harmonic mean (as a unified evaluation).

including cartoonish and pixelated images, which are outside REFVNLI's training distribution of real-world video frames. Similarly, on ImagenHub (Table 2), REFVNLI matches or exceeds all baselines in both single- and multi-subject settings, with 5.9-point gains over the strongest non-finetuned model on *subject preservation* of the multi-subject setting (Crop-IR). Lastly, on KITTEN (Table 3), REFVNLI leads in *textual alignment* but underperforms in *subject preservation*, though it remains statistically comparable to most baselines. This may result from REFVNLI's identity-sensitive training, which penalizes minor deviations—especially challenging for landmarks with intricate visual details (Fig. 12)—and from a domain shift, as landmarks were absent from REFVNLI's training data (OOD).

Notably, across all benchmarks, fine-tuning only for *textual alignment* (PaliGemma$_{\text{text}}$) slightly reduces performance, especially for *Animals* and *Humans*, while training solely for *subject preservation* (PaliGemma$_{\text{ref}}$) yields even larger declines—up to a 14.8 points for *Humans* (Table 1). This suggests that joint training provides complementary benefits, with subject preservation gaining the most.

Fig. 2 further showcases REFVNLI's strengths, like its sensitivity to subtle *textual alignment* errors, such as a missing waterfall (top-left). For *subject preservation*, it remains robust to identity-agnostic changes, like a zoomed-out parrot or person (top-center and bottom-center) or different clothes (bottom-center), while staying sensitive to key identity traits, e.g., changed facial features (bottom-left) and colors (left and middle balloons).

Overall, REFVNLI consistently outperforms or statistically matches all baselines on both criteria, with the only exception of *subject preservation* in the OOD landmarks category, where it still performs competitively. Importantly, it offers the best trade-off between *textual alignment* and *subject preservation*, surpassing all non-finetuned metrics in *Unified Evaluation* across all benchmarks.

5 Applicability to Rare Entities

To test REFVNLI on unfamiliar subjects, we use the ImageRAG benchmark (Shalev-Arkushin et al., 2025), which evaluates generated images based on prompts and reference images of uncommon subjects (e.g., scientific animal names, lesser-known dishes). Human annotators compared image pairs, selecting the better one based on *Textual Alignment*, *Visual Quality* (evaluating general depiction of the *entity* rather than exact reference-adherence), and *Overall Preference*. We report per-axis accuracy, defined as the frequency with which a metric ranks the human-preferred image higher, with approximate ties handled by rounding. *Overall Preference* is computed as the harmonic mean of textual and visual scores, and significance testing is assessed via bootstrap resampling, akin to §3.1.

	Textual Alignment	Visual Quality	Overall Preference
CLIP	51.8↓	91.3	69.2↓
DINO	-	91.4	-
Crop-IR	-	86.4↓	-
CLIPScore	47.4↓	-	-
BLIPScore	39.6↓	-	-
SigLIP	74.8↓	-	-
VQAScore	52.3↓	-	-
VIEScore	60.8↓	65.4↓	69.6↓
DreamBench++	56.6↓	83.1↓	78.9↓
PaliGemma$_{\text{text/ref}}$	61.6↓	83.0↓	83.0↓
REFVNLI	**87.2**	**95.5**	**91.4**

Table 4: **Results on ImageRAG rare concepts**, where users select the better image in each pair based on textual alignment, visual quality (general entity depiction rather than specific subject adherence), and overall preference. We report accuracy: how often models ranked the human-preferred image higher. Overall preference is the harmonic mean of textual and visual scores.

As shown in Table 4, REFVNLI consistently outperforms all baselines in aligning with human preferences across these criteria, showcasing strong robustness to rare subjects.[7] This is further supported by Fig. 6, where only REFVNLI repeatedly matches human selections.

6 Related Work

Evaluation of Visual Language Models (VLMs) spans various settings, including visual reasoning (Bitton-Guetta et al., 2024; Kahou et al., 2017) and visual question-answering (Antol et al., 2015; Marino et al., 2019; Mensink et al., 2023). For text-to-image (T2I) models, assessments normally focus on image quality (Heusel et al., 2017; Salimans et al., 2016), diversity (Rassin et al., 2024), and alignment with the text (Hessel et al., 2021; Radford et al., 2021; Hu et al., 2023; Yarom et al., 2023; Zhai et al., 2023; Lin et al., 2024). Assessing subject preservation, which is crucial for subject-driven generation, is typically done using embedding-based metrics like CLIP (Radford et al., 2021) and DINO (Caron et al., 2021). Other metrics, like VIEScore (Ku et al., 2024a) and DreamBench++ (Peng et al., 2024), use GPT-4o (OpenAI, 2024) to measure both criteria.

Subject-driven T2I models have been gaining much traction, with some methods fine-tuning general models into specialist versions that capture specific subjects and styles (Gal et al., 2022; Kumari et al., 2023; Ruiz et al., 2023; Sohn et al., 2023; Park et al., 2024). Others focus on broader applicability using one-shot examples, either through adapter-based methods that integrate encoded reference images into diffusion models (Gal et al., 2023; Jia et al., 2023; Wei et al., 2023; Ye et al., 2023) or via adapter-free techniques that directly use extracted features such as attention maps (Liu et al., 2023; Hertz et al., 2024; Lv et al., 2024).

Closely related, image editing complements subject-driven T2I generation in that the generated image's appearance is primarily governed by the input image, with the text only impacting specific aspects, whereas in out setting it is the other way around. The task has evolved from pixel-to-pixel translation for predefined transformations (Isola et al., 2017; Zhu et al., 2017; Wang et al., 2018) to more flexible, text-guided edits (Brooks et al., 2023; Tumanyan et al., 2023; Parmar et al., 2023), with recent diffusion-based methods improving precision via cross-attention manipulation (Hertz et al., 2022; Yang et al., 2023). Beyond images, personalized generation extends to other modalities, including videos and texts. Video generation can be conditioned on text (Li et al., 2018; Hong et al., 2022; Singer et al., 2022), reference images (Wei et al., 2024; Zhou et al., 2024), or other videos (Ku et al., 2024c). In text generation, efforts focus on style transfer (Reif et al., 2022; Zhang et al., 2024), debiasing (Zhao et al., 2018; Ravfogel et al., 2020), and broader semantic control (Shapira et al., 2022; Slobodkin et al., 2023; Xie et al., 2023).

Finally, several studies leveraged intra-frame relationships in videos to learn more human-aligned visual representations. These works aim to improve robustness to identity-agnostic variations (e.g., rotation, lighting), by analyzing consecutive frames sourced from public video datasets (Jin et al., 2018; Parthasarathy et al., 2023; Wang and Gupta, 2015; Wang et al., 2017; Wu and Wang, 2021) or captured by cameras on moving agents (Agrawal et al., 2015; Jayaraman and Grauman, 2015).

7 Conclusion

We present REFVNLI, a cost-effective and reliable metric for subject-driven T2I evaluation that jointly assesses *textual alignment* and *subject preservation*. Trained on a large-scale, auto-generated dataset, REFVNLI is designed to be robust to identity-agnostic visual variations (e.g., pose, lighting, background) while remaining sensitive to identity-

[7] TIFA was excluded due to assigning identical scores to 61% of pairs, making accuracy calculations unreliable.

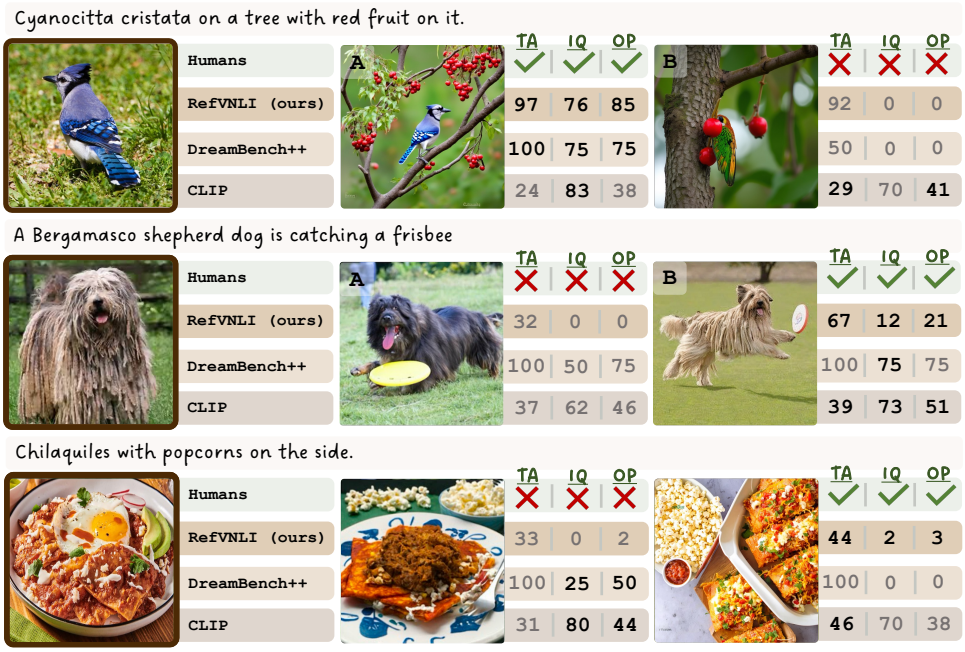

Figure 6: **ImageRAG Rare Entities Examples:** We compare REFVNLI with CLIP and DreamBench++ in aligning with human preferences (top rows of each example) across **Textual Alignment (TA)**, **Image Quality (IQ)**, and **Overall Preference (OP)**. DreamBench++ scores (0–4) are rescaled to 0–100 for readability. The higher of the two criterion-wise scores is emphasized unless both are equal.

specific features (e.g., facial features, object shape, and unique details) when evaluating subject preservation. For textual alignment, it leverages subject-specific prompts with perturbed hard negatives to detect and penalize fine-grained mismatches. Across benchmarks, REFVNLI outperforms or rivals all baselines, including larger GPT-4o-based metrics, particularly on less-common subjects.

Future work should focus on improving performance across artistic styles as well as identity-altering edits, and supporting multiple reference images. More broadly, REFVNLI facilitates progress in personalized T2I generation by enabling better checkpoint selection, reinforcement learning, and iterative model refinement.

8 Limitations

REFVNLI was trained on data sourced from real-life video frames and images. While it performs well on stylistically consistent inputs, including cartoonish or pixelated images, it struggles with cross-style scenarios where $image_{ref}$ and $image_{tgt}$ differ in style, as well as when the subject undergoes explicit modifications (e.g., changes in color or shape). Additionally, the current framework is limited to single-reference cases and should be extended to support multiple references, both for the same subject and for distinct ones.

Moreover, research on subject-driven generation could benefit from a unified score capturing overall performance, rather than the two separate scores currently provided by REFVNLI. Although the harmonic mean of the two offers a reasonable proxy, future iterations should aim to output a single, integrated metric, alongside the individual, more granular scores.

References

Pulkit Agrawal, Joao Carreira, and Jitendra Malik. 2015. Learning to see by moving. In *Proceedings of the IEEE international conference on computer vision*, pages 37–45.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, and 1 others. 2024. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*.

Nitzan Bitton-Guetta, Aviv Slobodkin, Aviya Maimon, Eliya Habba, Royi Rassin, Yonatan Bitton, Idan Szpektor, Amir Globerson, and Yuval Elovici. 2024.

Visual riddles: a commonsense and world knowledge challenge for large vision and language models. *arXiv preprint arXiv:2407.19474*.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660.

Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W Cohen. 2024. Subject-driven text-to-image generation via apprenticeship learning. *Advances in Neural Information Processing Systems*, 36.

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

Bradley Efron. 1987. Better bootstrap confidence intervals. *Journal of the American statistical Association*, 82(397):171–185.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2023. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM Transactions on Graphics (TOG)*, 42(4):1–13.

Brian Gordon, Yonatan Bitton, Yonatan Shafir, Roopal Garg, Xi Chen, Dani Lischinski, Daniel Cohen-Or, and Idan Szpektor. 2024. Mismatch quest: Visual and textual feedback for image-text misalignment. In *European Conference on Computer Vision*, pages 310–328. Springer.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. 2024. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4775–4785.

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.

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. *Advances in neural information processing systems*, 30.

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

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. True: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417.

Hsin-Ping Huang, Xinyi Wang, Yonatan Bitton, Hagai Taitelbaum, Gaurav Singh Tomar, Ming-Wei Chang, Xuhui Jia, Kelvin C. K. Chan, Hexiang Hu, Yu-Chuan Su, and Ming-Hsuan Yang. 2024. Kitten: A knowledge-intensive evaluation of image generation on visual entities. *Preprint*, arXiv:2410.11824.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134.

Dinesh Jayaraman and Kristen Grauman. 2015. Learning image representations tied to ego-motion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1413–1421.

Xuhui Jia, Yang Zhao, Kelvin CK Chan, Yandong Li, Han Zhang, Boqing Gong, Tingbo Hou, Huisheng Wang, and Yu-Chuan Su. 2023. Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. *arXiv preprint arXiv:2304.02642*.

SouYoung Jin, Aruni RoyChowdhury, Huaizu Jiang, Ashish Singh, Aditya Prasad, Deep Chakraborty, and Erik Learned-Miller. 2018. Unsupervised hard example mining from videos for improved object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 307–324.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2017. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*.

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. 2024a. VIEScore: Towards explainable metrics for conditional image synthesis evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12268–12290, Bangkok, Thailand. Association for Computational Linguistics.

Max Ku, Tianle Li, Kai Zhang, Yujie Lu, Xingyu Fu, Wenwen Zhuang, and Wenhu Chen. 2024b. Imagenhub: Standardizing the evaluation of conditional image generation models, 2024. *URL https://arxiv.org/abs/2310.01596*.

Max Ku, Cong Wei, Weiming Ren, Harry Yang, and Wenhu Chen. 2024c. Anyv2v: A tuning-free framework for any video-to-video editing tasks. *arXiv preprint arXiv:2403.14468*.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1931–1941.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, and 1 others. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981.

Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. 2020. TVQA+: Spatio-temporal grounding for video question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8211–8225, Online. Association for Computational Linguistics.

Dongxu Li, Junnan Li, and Steven Hoi. 2024. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Yitong Li, Martin Min, Dinghan Shen, David Carlson, and Lawrence Carin. 2018. Video generation from text. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2024. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pages 366–384. Springer.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. 2024b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *Preprint*, arXiv:2303.05499.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. 2024c. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22139–22149.

Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. 2023. Cones: concept neurons in diffusion models for customized generation. In *Proceedings of the 40th International Conference on Machine Learning*, pages 21548–21566.

Henglei Lv, Jiayu Xiao, and Liang Li. 2024. Pick-and-draw: Training-free semantic guidance for text-to-image personalization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10535–10543.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

Thomas Mensink, J Uijlings, Lluís Castrejon, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, A Araujo, and Vittorio Ferrari. 2023. Encyclopedic vqa: Visual questions about detailed properties of fine-grained categories. in 2023 ieee. In *CVF International Conference on Computer Vision (ICCV)*, pages 3090–3101.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, and 1 others. 2022. Simple open-vocabulary object detection. In *European Conference on Computer Vision*, pages 728–755. Springer.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Jae Wan Park, Sang Hyun Park, Jun Young Koh, Junha Lee, and Min Song. 2024. Cat: Contrastive adapter training for personalized image generation. *arXiv preprint arXiv:2404.07554*.

Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. 2023. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 conference proceedings*, pages 1–11.

Nikhil Parthasarathy, SM Eslami, Joao Carreira, and Olivier Henaff. 2023. Self-supervised video pretraining yields robust and more human-aligned visual representations. *Advances in Neural Information Processing Systems*, 36:65743–65765.

Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. 2024. Dreambench++: A human-aligned benchmark for personalized image generation. *Preprint*, arXiv:2406.16855.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Royi Rassin, Aviv Slobodkin, Shauli Ravfogel, Yanai Elazar, and Yoav Goldberg. 2024. Grade: Quantifying sample diversity in text-to-image models. *arXiv preprint arXiv:2410.22592*.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256.

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. A recipe for arbitrary text style transfer with large language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 837–848.

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.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *Advances in neural information processing systems*, 29.

Rotem Shalev-Arkushin, Rinon Gal, Amit H Bermano, and Ohad Fried. 2025. Imagerag: Dynamic image retrieval for reference-guided image generation. *arXiv preprint arXiv:2502.09411*.

Ori Shapira, Ramakanth Pasunuru, Mohit Bansal, Ido Dagan, and Yael Amsterdamer. 2022. Interactive query-assisted summarization via deep reinforcement learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2551–2568, Seattle, United States. Association for Computational Linguistics.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, and 1 others. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*.

Aviv Slobodkin, Niv Nachum, Shmuel Amar, Ori Shapira, and Ido Dagan. 2023. SummHelper: Collaborative human-computer summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 554–565, Singapore. Association for Computational Linguistics.

Kihyuk Sohn, Lu Jiang, Jarred Barber, Kimin Lee, Nataniel Ruiz, Dilip Krishnan, Huiwen Chang, Yuanzhen Li, Irfan Essa, Michael Rubinstein, and 1 others. 2023. Styledrop: Text-to-image synthesis of any style. *Advances in Neural Information Processing Systems*, 36:66860–66889.

Sahar Tahmasebi, Eric Müller-Budack, and Ralph Ewerth. 2025. Verifying cross-modal entity consistency in news using vision-language models. In *European Conference on Information Retrieval*, pages 339–354. Springer.

Gemini Team. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. 2023. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807.

Xiaolong Wang and Abhinav Gupta. 2015. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802.

Xiaolong Wang, Kaiming He, and Abhinav Gupta. 2017. Transitive invariance for self-supervised visual representation learning. In *Proceedings of the IEEE international conference on computer vision*, pages 1329–1338.

Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Fuxiao Liu, Gedas Bertasius, Mohit Bansal, Huaxiu Yao, and Furong Huang. 2024. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 416–442, Bangkok, Thailand. Association for Computational Linguistics.

Yujie Wei, Shiwei Zhang, Hangjie Yuan, Xiang Wang, Haonan Qiu, Rui Zhao, Yutong Feng, Feng Liu, Zhizhong Huang, Jiaxin Ye, and 1 others. 2024. Dreamvideo-2: Zero-shot subject-driven video customization with precise motion control. *arXiv preprint arXiv:2410.13830*.

Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. 2023. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15943–15953.

Daniel Winter, Asaf Shul, Matan Cohen, Dana Berman, Yael Pritch, Alex Rav-Acha, and Yedid Hoshen. 2024. Objectmate: A recurrence prior for object insertion and subject-driven generation. *arXiv preprint arXiv:2412.08645*.

Haiping Wu and Xiaolong Wang. 2021. Contrastive learning of image representations with cross-video cycle-consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10149–10159.

Yujia Xie, Xun Wang, Si-Qing Chen, Wayne Xiong, and Pengcheng He. 2023. Interactive editing for text summarization. *arXiv preprint arXiv:2306.03067*.

Fei Yang, Shiqi Yang, Muhammad Atif Butt, Joost van de Weijer, and 1 others. 2023. Dynamic prompt learning: Addressing cross-attention leakage for text-based image editing. *Advances in Neural Information Processing Systems*, 36:26291–26303.

Michal Yarom, Yonatan Bitton, Soravit Changpinyo, Roee Aharoni, Jonathan Herzig, Oran Lang, Eran Ofek, and Idan Szpektor. 2023. What you see is what you read? improving text-image alignment evaluation. *Advances in Neural Information Processing Systems*, 36:1601–1619.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. Alignscore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986.

Chiyu Zhang, Honglong Cai, Yuexin Wu, Le Hou, Muhammad Abdul-Mageed, and 1 others. 2024. Distilling text style transfer with self-explanation from llms. *arXiv preprint arXiv:2403.01106*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.

Yufan Zhou, Ruiyi Zhang, Jiuxiang Gu, Nanxuan Zhao, Jing Shi, and Tong Sun. 2024. Sugar: Subject-driven video customization in a zero-shot manner. *arXiv preprint arXiv:2412.10533*.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.

A Reproducibility and Model Selection

We fine-tuned PaliGemma (Beyer et al., 2024) on a balanced subset of our dataset, created by undersampling the more frequent labels. Training was conducted over 24 hours using two NVIDIA A100 GPUs (80 GiB each) with a batch size of 4. During training, we enclosed the referenced subject in the *prompt* with <*u*> and <*u*> markups, and provided it alongside the separately passed $image_{ref}$ and $image_{tgt}$. The model was trained to generate one of four strings—*'00'*, *'01'*, *'10'*, *'11'*—where the first and second digits represent *textual alignment* and *subject preservation*, respectively. See Fig. 7 for an input-output example.

At inference, we used the same input structure, including the insertion of markups around the subject in the *prompt* and separate image inputs. The *textual alignment* and *subject preservation* scores were then computed from the probabilities of the first and second generated tokens, respectively, with each score defined as the probability of token *'1'* divided by the sum of probabilities for tokens *'0'* and *'1'*.

Additionally, we explored fine-tuning LLaVA-1.5 (Liu et al., 2024a)[8] as an alternative backbone model, and determined that PaliGemma achieved a strongest performance on our development set.

B Additional Details on the Meta-evaluation

For meta-evaluation, we perform bootstrapping (Efron, 1987) (1,000 samples per benchmark, with repetitions). For the ImageRAG rare entities benchmark, consisting of only 26 instances, we sample each time a sample 4 times this size, to ensure there are at least 100 instances. For the other benchmarks, which are significantly larger, the size of each sample is identical to the original size of the corresponding benchmark. For the calculations of confidence intervals (CIs), we use significance level $p = 0.05$.

C Further Information on the Data Construction Pipeline

C.1 Collection of Subject-driven Image Pairs

To reduce noise when collecting subject-driven image pairs, we applied several filtering steps, including the removal of blurred images and those not depicting the intended subject. Subject presence was verified using Gemini (Team, 2024) (version *gemini-2.0-flash*). For the subset sourced from TVQA+, we additionally filtered out frames containing subtitles or credits, also using Gemini.

For identity-sensitive image pairs, we used Stable Diffusion (Rombach et al., 2022) for inpainting,[9] with $\eta = 1.0$ and a guidance scale of 3.0. We retained only images with a full mask size of at least 60,000 pixels (20,000 for humans, focusing on facial regions). Five patches of 250–300 pixels were randomly sampled and inpainted. To increase the likelihood of meaningful subject changes, we further filtered out inpainted images where all patchwise MSE values fell below 6,500 for objects, 5,400 for animals, and 20,000 for humans.

C.2 Collection of Image-*prompt* Pairs

For the image-captioning of the $image_{tgt}$ with the inserted bounding boxes, we employed Gemini (Team, 2024) (version *gemini-2.0-flash*).

C.3 Generation of *prompt-*$image_{tgt}$ Hard-negatives

Fig. 8 showcases the prompt used to generated the *prompt-*$image_{tgt}$ hard negatives.

D Ablations

To assess the impact of various design decisions in REFVNLI, we run an ablation study examining alternative input and output configurations. On the output side, we test: reversing the classification order (subject preservation before textual alignment); a 4-label multiclass framework for joint text-image alignment classification; and a model that prefixes a designated token (*'TEXT'* or *'IMAGE'*) to the *prompt* to enable separate classification of each aspect within a unified model. For inputs, we explore the effect of removing subject markup from the *prompt* and of concatenating $image_{ref}$ and $image_{tgt}$ instead of passing them separately. Finally, we also explore the impact of omitting the identity-sensitive training examples (see §2.1), by only using the video-based instances during training.

Results (Table 5) show that reversing the classification order degrades performance, particularly in subject preservation, as does evaluating each aspect separately. This suggests that first evaluating textual alignment helps in subject preservation

[8] https://huggingface.co/llava-hf/llava-1.5-7b-hf

[9] https://huggingface.co/stabilityai/stable-diffusion-2-inpainting

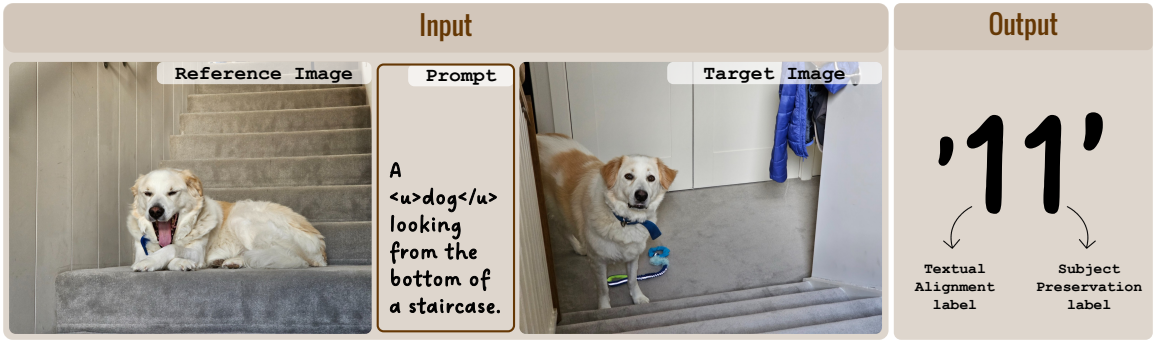

Figure 7: Example of the input and output of REFVNLI. The input consists of $image_{ref}$ and $image_{tgt}$, as well as from the *prompt* with <u> and <\u> markups around the referenced subject (e.g., dog), while the output consists of two digits of *'0'* or *'1'*, with the first digit representing the first and second digits being the *textual alignment* and *subject preservation* labels, respectively.

	Textual Alignment			Subject Preservation			Unified Evaluation		
	DreamBench++	ImagenHub (Single/Multi)	KITTEN	DreamBench++	ImagenHub (Single/Multi)	KITTEN	DreamBench++	ImagenHub (Single/Multi)	KITTEN
REFVNLI (ours)	81.5	87.7 / 86.3	97.0	82.7	83.0 / 62.8	82.3	82.1	85.3 / 72.7	89.0
reverse classification order	80.0	85.2 / 85.5	95.3	80.9	84.3 / 68.7	87.0	80.4	84.7 / 76.2	91.0
multiclass	79.5	83.7 / 84.7	94.7	79.6	76.0 / 61.1	86.3	79.5	79.7 / 71.0	90.3
separate classification	79.7	85.2 / 87.5	95.8	78.3	77.1 / 56.7	89.2	79.0	80.9 / 68.8	92.4
no markup	78.4	87.0 / 84.3	92.3	65.5	75.9 / 60.8	88.7	71.4	81.1 / 70.6	90.5
concatenated images	79.6	86.2 / 86.2	93.6	74.2	81.1 / 81.2	89.9	76.8	83.6 / 83.6	91.7
only video-based training	80.4	85.2 / 83.8	96.3	77.6	77.5 / 66.4	89.7	79.0	81.2 / 74.1	92.9

Table 5: **Ablation Study:** ROC AUC scores for various ablated versions of REFVNLI across benchmarks (over all subjects). The ablations evaluate alternative output formulations, such as reversed classification order, a four-label multiclass framework, and separate aspect classification via a special token. Additional variants exclude subject markup from input prompt, merge reference and target images, or remove identity-sensitive training examples by using only video-based instances.

assessment. The multiclass approach also underperforms compared to our dual binary classification setup, highlighting the benefits of treating each criterion independently. Further, removing subject markup weakens subject preservation, underscoring its role in linking the reference image to the *prompt*. Additionally, concatenating images instead of processing them separately harms performance, emphasizing the advantage of distinct image inputs. Finally, excluding identity-sensitive training instances leads to notable drops in subject preservation, underscoring their importance.

	DreamBench++			ImagenHub			KITTEN
	Animal	Human	Object	Animal	Object	Multi-subj.	Landmarks
OWL-ViT	77.4	**87.1**	81.8	74.8	83.1	52.0	85.6
GroundingDINO	**76.9**	85.6	**83.4**	**77.6**	**84.1**	**56.8**	**90.2**

Table 6: **ROC AUC scores of Crop-IR for subject preservation** when employed with two different object-detection models: OWL-ViT and GroundingDino. Bold indicates the highest score per column.

E Crop-IR Object Detection Model Ablation

Deploying the Crop-IR metric (Winter et al., 2024) requires an object detection model to locate and crop the referenced subjects. To this end, we compare two prominent object detection models: OWL-ViT (Minderer et al., 2022)[10] and GroundingDino (Liu et al., 2024b).[11] As shown in Table 6, GroundingDino leads to better evaluation of subject preservation, and is therefore adopted to ensure a fairer evaluation.

F Computational Cost

Table 7 presents the computational costs of all baseline models and REFVNLI, in terms of inference time, GPU memory usage and GPT-4o API costs.

[10] https://huggingface.co/google/owlv2-base-patch16

[11] https://huggingface.co/IDEA-Research/grounding-dino-base

Context

Misalignment Injection Instructions (Short Captions)

1. **Understand the Caption:** Carefully read the short caption to fully grasp the scene it describes.
2. **Identify and Swap:** Select a single visual detail within the caption to modify. Replace this detail with a *different, incorrect, but still plausible* visual detail. For example, you might change a color, an object, or a location. **Do not modify the underlined entity (if any).**
3. **Apply the Tags:** Enclose the *original* visual detail within `<swap>` tags. Immediately *after* the closing `</swap>` tag, write the *new, incorrect* visual detail. There should be no space between the closing `</swap>` and the new word.

   Example: If the original sentence is "The cat sat on the red mat," and you want to change "red" to "blue," the result should be: "The cat sat on the `<swap>red</swap><blue>` mat."
4. **Final Check:** Ensure the modified caption is grammatically correct and reads naturally, even though it now contains a factual error. The sentence should be internally logical, despite contradicting the actual visual content. Again, **ensure the underlined entity (if any) remains completely unchanged.**

Few-Shot

Here are some examples:

INPUT: A woman is sitting in a living room, and `<u>she</u>` is looking at something with a concerned expression
OUTPUT: A woman is sitting in a `</swap>`living room`</swap><kitchen>`, and `<u>she</u>` is looking at something with a concerned expression.

INPUT: Two men are sitting on a leather couch in a living room. One `<u>man</u>` is sitting on the left side of the couch, looking at a laptop. The other man is sitting on the right side of the couch, talking on a phone. The room is decorated with various items, including a large model of a spaceship.
OUTPUT: Two men are sitting on a leather couch in a living room. One `<u>man</u>` is sitting on the left side of the couch, looking at a laptop. The other man is sitting on the right side of the couch, talking on a phone. The room is decorated with various items, including a large model of a `<swap>spaceship</swap><sailboat>`.

Now it's your turn! Follow the instructions. Answer only with the corrupted sentence, Don't forget to add the tags.

INPUT: A lizard is perched on a rock, surrounded by other rocks and foliage. The `<u>lizard</u>` is facing the camera, with its head raised and its tail curled behind it.

OUTPUT:

Generated

OUTPUT: A lizard is perched on a **`<swap>rock</swap><branch>`**, surrounded by other rocks and foliage. The `<u>lizard</u>` is facing the camera, with its head raised and its tail curled behind it.

Figure 8: Hard Negative Caption Generation. This figure illustrates the prompting strategy used to generate hard negative captions, containing a single, plausible but factually incorrect visual detail, for enhanced misalignment detection.

	Inference Time (seconds)	GPU Memory Usage (GiB)	API Calls Cost ($)
CLIP	0.1	1.2	-
DINO	0.06	0.7	-
Crop-IR	0.6	5.8	-
ArcFace	1.2	-	-
CLIPScore	0.07	0.6	-
BLIPScore	0.7	4.5	-
SigLIP	0.03	1.4	-
TIFA	22.5	26.4	-
VQAScore	0.2	23.1	-
VIEScore	6.9	-	0.04
DreamBench++ (text)	3.2	-	0.02
DreamBench++ (ref)	1.0	-	0.03
PaliGemma$_{\text{text}}$	0.4	12.5	-
PaliGemma$_{\text{ref}}$	0.4	12.5	-
REFVNLI	0.5	12.5	-

Table 7: Computational costs for all baseline models and REFVNLI, including per-instance inference time (in seconds), GPU memory usage (in GiB), and GPT-4o API costs (in $, only when applicable), averaged across benchmarks. For DreamBench++, we report separate values for each evaluation criterion, as each requires a distinct API call under its framework. The final three rows present models fine-tuned on our dataset, with PaliGemma$_{\text{text}}$ and PaliGemma$_{\text{ref}}$ being the variants tuned exclusively for evaluating textual alignment and subject preservation, respectively.

G Additional Qualitative Examples for Subject Preservation Evaluation

In Figures 9, 10, 11, and 12 we present additional qualitative examples of *subject preservation* evaluation for the *Animal*, *Human*, *Object*, and *Landmark* categories, respectively.

H List of Data and Software Licenses Employed in this Paper

Our framework dependencies are:

1. Mementos dataset: https://github.com/si0wang/Mementos, Misc.
2. TVQA+ dataset: https://github.com/jayleicn/TVQAplus/blob/master/LICENSE, under the MIT License.
3. Open Images dataset: https://github.com/openimages/dataset/blob/main/LICENSE, under an Apache License 2.0.
4. PaliGemma model: https://ai.google.dev/gemma/terms, under Gemma Terms of Use License.
5. Gemini model: https://ai.google.dev/gemini-api/docs/models, under an Apache License 2.0.
6. GPT-4o model: https://github.com/openai/openai-openapi/blob/master/LICENSE, under the MIT License.

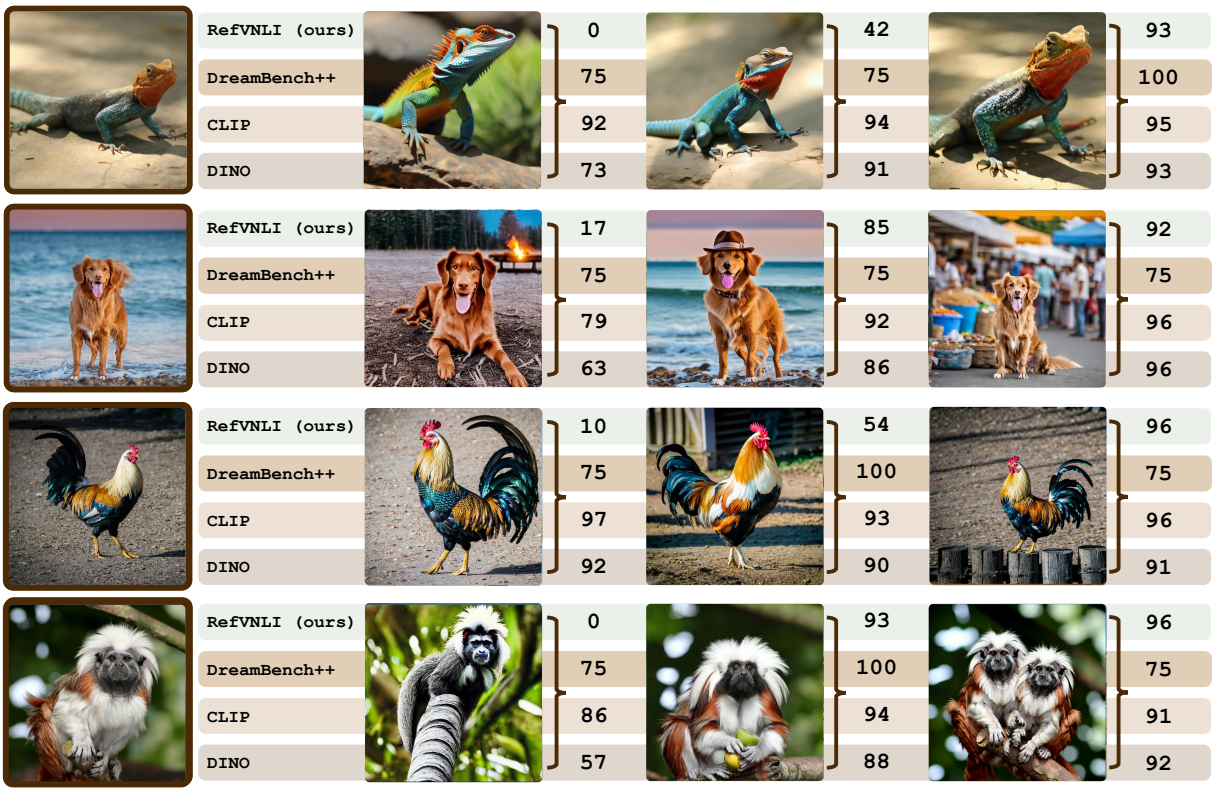

Figure 9: **Qualitative Examples of Subject Preservation Evaluation for the *Animal* Category.** DreamBench++ scores (0-4) are scaled to 0-100 for better readability.

7. OWL-ViT model: https://huggingface.co/google/owlv2-base-patch16, under an Apache License 2.0.

8. GroundingDino model: https://huggingface.co/IDEA-Research/grounding-dino-base, under an Apache License 2.0.

9. Stable Diffusion inpainting model: https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0/blob/main/LICENSE.md, under Stability AI CreativeML Open RAIL++-M License.

10. LLaVA-1.5 model: https://github.com/haotian-liu/LLaVA/blob/main/LICENSE, under an Apache License 2.0.

Figure 10: **Qualitative Examples of Subject Preservation Evaluation for the *Human* Category.** DreamBench++ scores (0-4) are scaled to 0-100 for better readability.

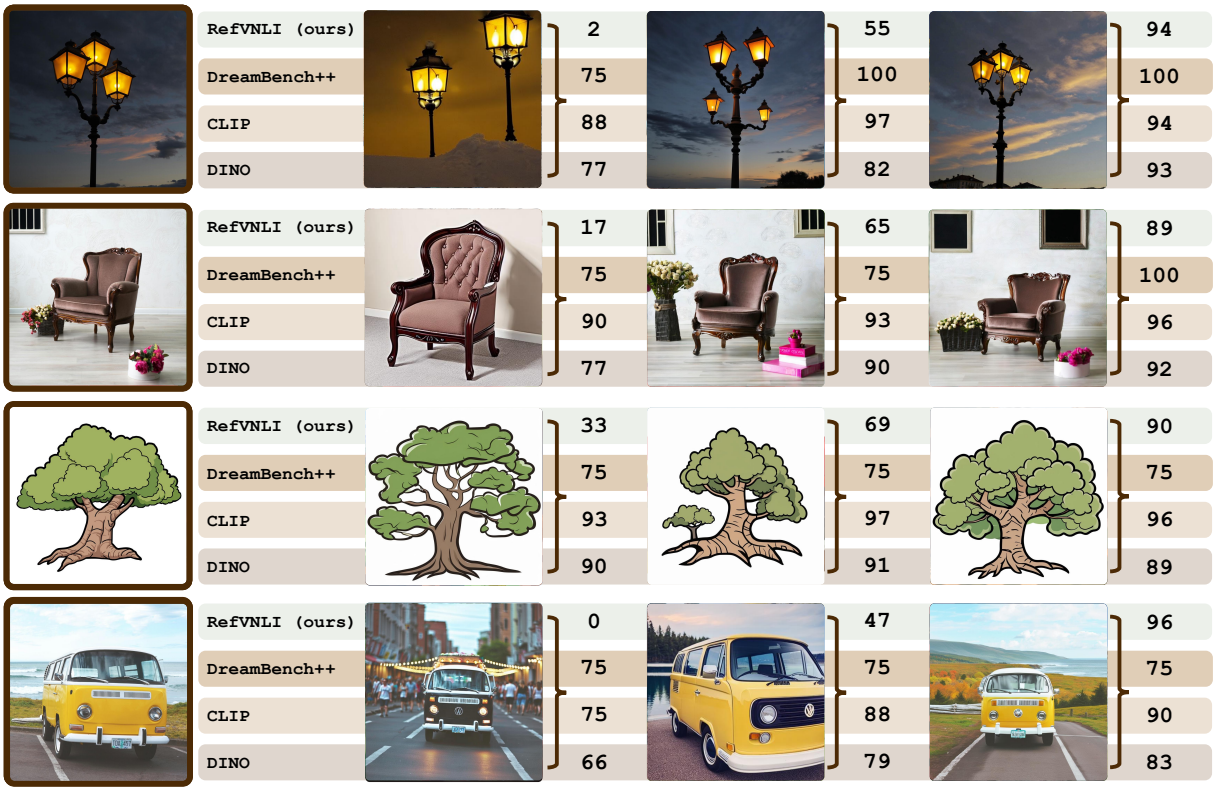

Figure 11: **Qualitative Examples of Subject Preservation Evaluation for the *Object* Category.** DreamBench++ scores (0-4) are scaled to 0-100 for better readability.

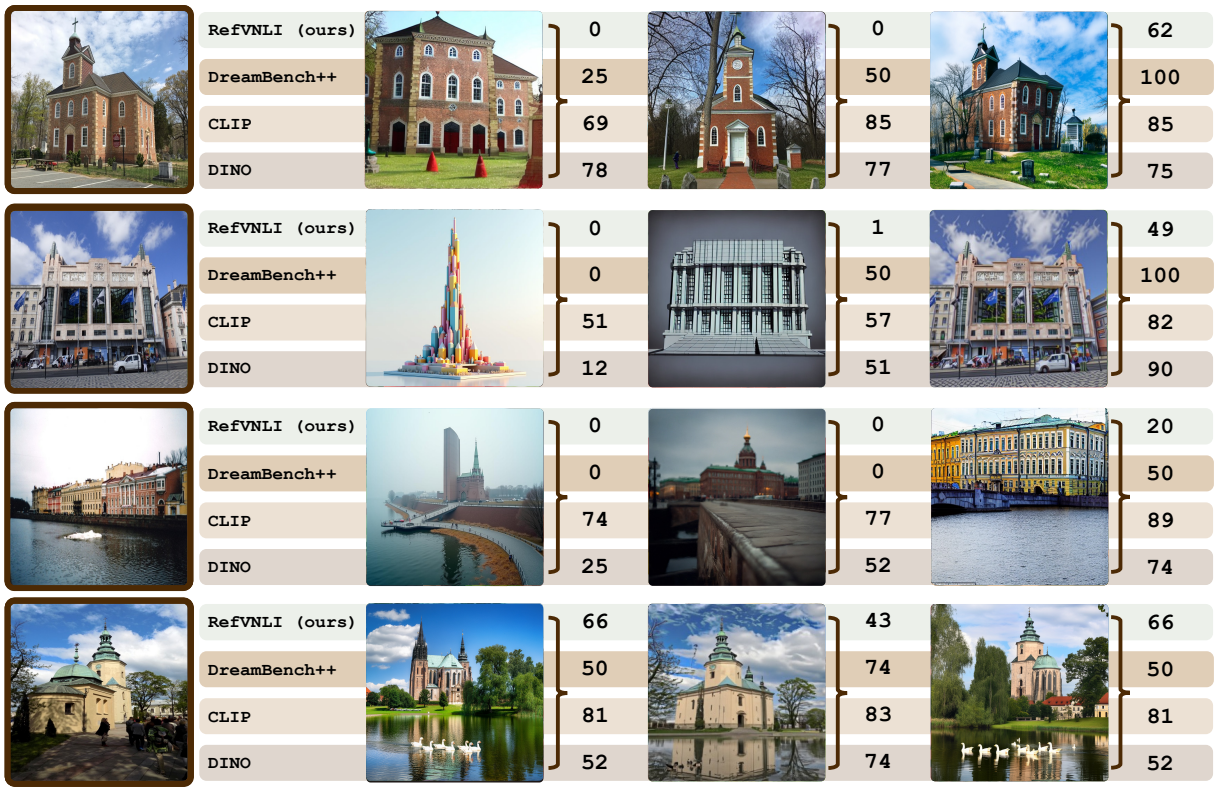

Figure 12: **Qualitative Examples of Subject Preservation Evaluation for the *Landmark* Category.** DreamBench++ scores (0-4) are scaled to 0-100 for better readability.