This is the peer reviewd version of the followng article:

Optimizing Resource Consumption in Diffusion Models through Hallucination Early Detection / Betti, F., Baraldi, L., Baraldi, L., Cucchiara, R., Sebe, N.. - (2024). (European Conference on Computer Vision Workshops Milan Sep 29th - Oct 4th).

Terms of use:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

15/10/2026 20:07

(Article begins on next page)

Optimizing Resource Consumption in Diffusion Models through Hallucination Early Detection

Federico Betti[1], Lorenzo Baraldi[2], Lorenzo Baraldi[3], Rita Cucchiara[3], and Nicu Sebe[1]

[1] University of Trento, Italy {federico.betti,nicu.sebe}@unitn.it
[2] University of Pisa, Italy lorenzo.baraldi@phd.unipi.it
[3] University of Modena and Reggio Emilia, Italy
{lorenzo.baraldi,rita.cucchiara}@unimore.it

Abstract. Diffusion models have significantly advanced generative AI, but they encounter difficulties when generating complex combinations of multiple objects. As the final result heavily depends on the initial seed, accurately ensuring the desired output can require multiple iterations of the generation process. This repetition not only leads to a waste of time but also increases energy consumption, echoing the challenges of efficiency and accuracy in complex generative tasks. To tackle this issue, we introduce HEaD (Hallucination Early Detection), a new paradigm designed to swiftly detect incorrect generations at the beginning of the diffusion process. The HEaD pipeline combines cross-attention maps with a new indicator, the Predicted Final Image, to forecast the final outcome by leveraging the information available at early stages of the generation process. We demonstrate that using HEaD saves computational resources and accelerates the generation process to get a *complete* image, *i.e.* an image where all requested objects are accurately depicted. Our findings reveal that HEaD can save up to 12% of the generation time on a two objects scenario and underscore the importance of early detection mechanisms in generative models.

1 Introduction

In the rapidly evolving domain of AI, generative models have emerged as a notable subfield, demonstrating an exceptional ability to generate complex visual and textual content [4, 19]. The advent of Text-to-Image (T2I) generation marked a significant leap in this domain through the introduction of GAN-based approaches [6, 23, 32] and further advancements through large-scale pre-trained Diffusion Models (DM) such as Stable Diffusion (SD) [24] and others [3, 21]. These approaches have been instrumental in shaping the generative AI landscape, delivering images that are increasingly indistinguishable from real ones.

Generative models, while progressing, often hallucinate "long-tail" objects, which are underrepresented elements in training datasets, and have significant shortcomings when generating multiple objects [3, 27, 33]. Furthermore, they

Fig. 1: Overview of the HEaD pipeline: during the generation process, HEaD assesses whether all designated objects will be accurately represented in the final image, determining if the generation process should continue or be restarted with a different seed.

frequently hallucinate attributes, counts, and semantic object relations, which is especially problematic when tasked with rendering scenes involving multiple objects [5, 14].

The challenge is further intensified when generating combinations of specific objects, where diffusion patterns often produce inconsistencies, significantly impacting the quality of the output [7]. The choice of the initial seed, which dictates the initial latent noise, is fundamental in navigating the latent space for correct image generation. The dependency on seed selection highlights the unpredictability and variability of these models [3, 12, 14, 26, 30]. Although automatic evaluation mechanisms could offer a potential solution to these challenges, their adoption is not straightforward. While some attempts in this direction have been made [2, 15], indeed, they still fail at ensuring a sufficiently fast and reliable evaluation. However, employing these automatic evaluations tends to be slow and resource-intensive. This is largely due to the numerous incorrect results, which require images to be regenerated, thus escalating both time and resource costs.

Addressing these challenges, we introduce HEaD (Hallucination Early Detection), the first approach designed to enhance both the efficiency and accuracy of generative DMs. HEaD incorporates the use of cross-attention maps to examine the relationship between the prompt and the internal attention layers of the model, along with the *Predicted Final Image* (PFI) - a prediction of the expected outcome at intermediate stages of the generation process. The combination of PFIs and cross-attention maps allows for the early identification of potential errors by predicting the presence of objects requested by the initial prompt. By preemptively detecting these anomalies, HEaD hints at stopping the generation diffusion process, thereby conserving resources and reducing the time spent on generating images that would not eventually meet quality standards. Aiming for a *complete* generation – where all requested objects are accurately depicted – halting the generation process based on a prediction of the final outcome proves to be far more efficient than completing an image generation and subsequently evaluating it. This approach not only streamlines the generation process but also enhances resource utilization by avoiding the production and evaluation of substandard images.

We trained two types of networks, each with a different backbone for handling PFI data, followed by CNN-based processing of cross-attention maps. This training occurred at different points in the generation pipeline to assess their

impact on prediction quality and potential time savings. Results indicate that using a Visual Transformer as a backbone yields superior outcomes. Moreover, while networks trained towards the later stages of the generation pipeline benefit from higher-quality input and thus demonstrate better performance, those trained earlier exhibit greater potential for time and resource savings. It is also important to note that the methodologies and models described in this study are model-agnostic, *i.e.* they can be seamlessly adapted to any diffusion-based generative model.

In this work, we focus on specific hallucinations: the omission in the generated image of one or more target objects indicated in the textual prompt. We propose both a detector (trained on a dataset of corrected and hallucinated generated images) and a general approach for time-saving prediction that accounts for both the hallucination probability of the specific generative model and the accuracy of the detector. For instance, when generating images with prompts involving two objects in non-trivial combinations, SD2 produces hallucinations or missing object errors in 41% of cases, according to our dataset. Our HEaD approach can detect the majority of these errors with minimal time overhead, thus saving up to 12% of the average generation time in this simple scenario.

To sum up, our main contributions are as follows:

- We introduce a new element, PFI, and demonstrate that its integration with cross-attention maps effectively facilitates the early detection of objects within generated images.
- We propose a comprehensive framework for time saving evaluation. We demonstrate the potential time and resource saving for the generation of *complete* images from multi-object prompts, without compromising the output integrity and generation quality.
- A novel classifier for Hallucination Early Detection has been developed that, when integrated into the diffusion process, combines information at each diffusion level and acts as an early evaluator. This classifier stops the process if a hallucination is detected, thereby enhancing the efficiency and accuracy of the generation.

2 Related Works

Text-to-Image Evaluation. Quantifying the alignment between the generated image and the initial prompt is a challenging task, and as of now, no effective solutions have been identified. Among the assessment metrics, CLIP-Score [9] evaluates the cosine similarity between the prompt and the image, both having undergone processing through their respective visual and textual CLIP backbones. Recent studies [2, 15] have proposed innovative scoring mechanisms that leverage the capabilities of Large Language Models (LLMs) and Visual Question Answering. In alignment with this research trajectory, various investigations [11, 31] have introduced diverse methodologies, positioning their work within the reasoning paradigm facilitated by LLMs.

Despite their success in identifying hallucinatory elements in generative models, these works still require the generated image as input, which is produced only in the final step of the diffusion process. Additionally, they incorporate evaluation steps beyond image generation, resulting in delays due to the reliance on foundational models within the evaluation pipeline. Conversely, HEaD enables the detection of hallucinations during the generative process itself, preventing the creation of images that do not align with their prompts.

Attention Maps in Image Generation. The integration of attention mechanisms has been a cornerstone in improving image synthesis within generative models. Notably, Chefer *et al.* refines these processes to enhance image detail [3]. Cross-attention layers [24] have significantly boosted visual fidelity, a concept further explored by Hertz *et al.* to maintain coherence between text prompts and visual outputs [8]. The importance of semantic layouts in improving image quality and interpretability has also been highlighted by Wand *et al.* [29]. Building on these ideas, SynGen was introduced [22], which aligns attention maps with prompt syntax to improve attribute correspondence, optimizing the generation process without the need for model retraining. Furthermore, Mao *et al.* developed a novel method for controlling image synthesis by editing initial noise images, demonstrating that manipulating pixel blocks in initial latent images can influence specific content generation [16]. Additionally, Balaji *et al.* proposed eDiff-I, an ensemble of expert denoisers that enhances text alignment and visual quality by specializing models for different stages of synthesis [1].

Following the consensus on the effectiveness of cross-attention as a telltale sign of the fidelity of the generation, our work exploits this information as a discriminating factor for the accurate generation of the final image.

Seed Importance. In Text-to-Image generation, images are significantly impacted by the initial state or starting seed of the diffusion process. Indeed, different seeds produce completely different image results as highlighted by Karthik *et al.*, which claims to generate better-aligned images by evaluating multiple seeds [12]. Furthermore, image editing by directly manipulating the initial noise instead of steering the generation process with additional mechanisms has also been proposed [16, 26].

Seed selection has gained relevance in the generation of long-tail concepts [33]. Samuel *et al.* propose that, in the generation of rare subjects, training predominantly involves exposure to a limited segment of the initial noisy latent space [27]. This selective exposure during training contributes to the generation of unsatisfactory outcomes across a majority of generative seeds at inference time. Hence, the exploration of diverse generative seeds remains a critical aspect in enhancing generative outcomes. To mitigate the occurrence of hallucinations, HEaD suggests altering the seed in the event of detecting hallucinations in the generative process.

3 Preliminaries

Latent Diffusion Models. We focus on the Stable Diffusion (SD) model [24], which leverages the latent space of an autoencoder rather than the conventional pixel image space. The process begins with an encoder E transforming an image x into a latent code $z = E(x)$. A decoder then D aims for accurate reconstruction, insisting $D(E(x)) \approx x$. Within this framework, a denoising diffusion probabilistic model (DDPM) [10,28] operates. This model works on the latent space, creating a denoised version of the input latent z_t at each timestep t. Notably, SD enhances this process by integrating a conditioning vector $c(y)$, typically derived from a textual prompt y through a CLIP text encoder [20]. The objective is to minimize the loss function:

$$L = \mathbb{E}_{z\sim E(x),y,\epsilon\sim\mathcal{N}(0,1),t} \left\|\epsilon - \epsilon_\theta(z_t, t, c(y))\right\|^2 \tag{1}$$

where ϵ_θ is a UNet network [25] with attention layers that aims to eliminate the added noise, considering the noisy latent z_t, timestep t, and conditioning encoding $c(y)$.

To obtain the final image from the denoised latent representation, the last step involves passing the final latent representation through a Variational Autoencoder (VAE) decoder. This decoder, denoted as D, translates the latent space back into the pixel space, thus completing the image generation process. The transition from the final latent state z_0 to the generated image x_0 can thus be described by

$$x_0 = D(z_0), \tag{2}$$

where D is the VAE Decoder [13] trained to map the latent representations to high-fidelity images. For further details, we refer the reader to [24].

Schedulers in Diffusion Models. In DMs, schedulers are employed to orchestrate the denoising steps, shaping the generation by modulating noise levels. These algorithms enable the transition from a noisy latent representation to a refined image without adversarial training. In our HEaD approach, we have tailored the scheduler's function to extract the PFI at intermediate stages. This modification aims to achieve the most accurate representation of the final image during the generation process.

The transition of latents z_t at time step t to another subsequent state $z_{t'}$ is described as follows. The predicted noise ϵ_t is firstly obtained from the output of the UNet model, and then the new latents $z_{t'}$ are computed through the update function of the scheduler Δ, as

$$\begin{aligned} \epsilon_t &= \epsilon_\theta(z_t, t) \\ z_{t'} &= \Delta(z_t, \epsilon_t, t, t'). \end{aligned} \tag{3}$$

Here, ϵ_t is informed by the current latents and time step, and Δ is the scheduler update function computing the new latents $z_{t'}$ based on the predicted noise ϵ_t. The behavior of Δ is determined by the specific scheduler chosen, which dictates the complex dynamics of the denoising process.

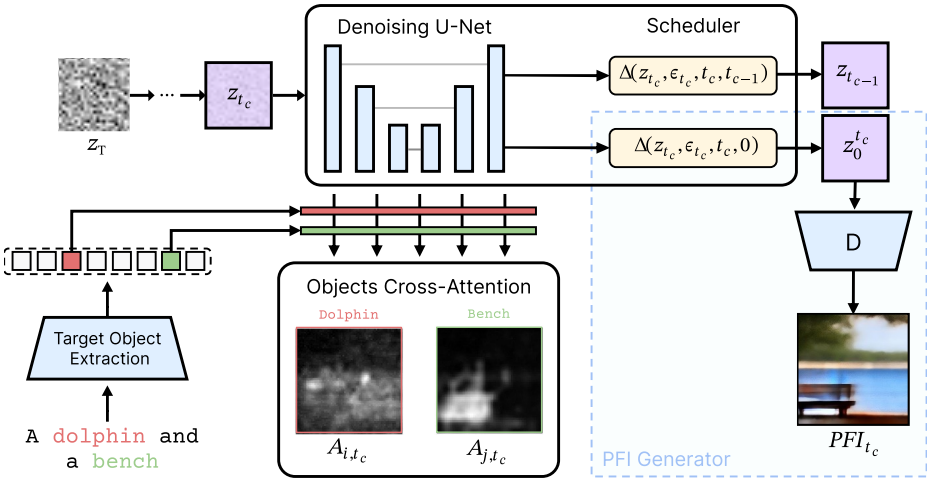

Fig. 2: This figure illustrates the process of extracting subjects, cross-attention maps and PFI at each critical timestep $t_c \in \mathcal{T}$. These elements serve as inputs for the HP network, which evaluates the presence of objects in the final image. For the depicted seed, the bench appears in the final image, whereas the dolphin does not.

4 Hallucination Early Detection

HEaD primary goal is to detect and preemptively interrupt faulty generative processes. Its novelty lies in its ability to perform this detection at intermediate stages of the image generation, leveraging one or more time steps of the diffusion pipeline. Consequently, if the Hallucination Prediction (HP) network predicts that the image will not be *complete* – indicating the absence of at least one target object – the generative process can be halted and restarted with an alternative seed. This preemptive detection conserves computational resources by preventing the completion of flawed images, eliminating the need to sample a new seed, and avoiding a complete restart from scratch.

In this section, we illustrate the proposed HEaD approach at inference time to streamline the generation process and, as a result, enable automatic quality assessment of the final output. The pipeline initial step involves extracting the target objects from the prompt and providing hallucination indicators for the HP network to evaluate.

4.1 Cross-Attention Maps and PFIs Extraction

Given a prompt y containing a set of target objects O to be generated, the extraction process of these target objects can be formalized as follows:

$$O = \text{TOE}(y) \tag{4}$$

where TOE$(\cdot)$ represents the Target Object Extraction function. Here, the term "objects" refers to words in the prompt directly associated with discernible elements in the image, for which we will extract the corresponding cross-attention

maps. While our current methodology primarily focuses on objects, it holds the capability for future expansion to include a wider spectrum of visual concepts, thereby transcending the confines of object-based extraction.

We define a sequence of *critical timesteps*, denoted as $\mathcal{T} = \{t_{c_1}, \ldots, t_{c_k}\}$, as specific steps in the diffusion process where cross-attention maps and PFI are extracted. These components will serve as inputs for the HP network.

In diffusion models, the UNet employs cross-attention layers at resolutions from 64 to 8, producing a combined attention map $A_t \in \mathbb{R}^{64\times64\times N}$, where N is the number of tokens from the prompt y. For each object o in the target set O and each critical timestep $t_c \in \mathcal{T}$, the specific cross-attention map A_{o,t_c} is derived by filtering A_t for object o.

For each critical timestep $t_c \in \mathcal{T}$, a Predicted Final Image (PFI_{t_c}) is extracted. PFI_{t_c} represents the prediction of the expected outcome at the end of the generation process, using only information available at timestep t_c. In particular, the scheduler projects the latents at t_c to the final step, and the decoder translates these predicted latents into the image space. The process is defined as follows:

$$\begin{aligned} \epsilon_{t_c} &= \epsilon_\theta(z_{t_c}, t_c) \\ z_0^{t_c} &= \Delta(z_{t_c}, \epsilon_{t_c}, t_c, 0) \\ \text{PFI}_{t_c} &= D(z_0^{t_c}) \end{aligned} \tag{5}$$

where ϵ_{t_c} represents the predictive noise obtained from the UNet model at critical timestep t_c. The function Δ updates the latents z_{t_c} to the predicted latents at the final timestep, denoted as $z_0^{t_c}$. Finally, the VAE decoder D translates these predicted final latents into PFI_{t_c}.

Examples of PFIs extracted at different timesteps are shown in Fig. 3. The collection of PFIs, namely $\text{PFI}_{\mathcal{T}}$, and attention maps, $A_{O,\mathcal{T}}$, across all critical timesteps provides a comprehensive dataset for the HP network to analyze and predict the presence of objects in the final image.

4.2 Hallucination Prediction Network

During the evaluation phase, the Hallucination Prediction network takes as input the cross-attention maps A_o for a specific object and the $\text{PFI}_{\mathcal{T}}$ and outputs a binary prediction indicating the presence or absence of that specific target object in the final image:

$$H_o = \text{HP}(A_{o,\mathcal{T}}, \text{PFI}_{\mathcal{T}}) \tag{6}$$

where H_o is the binary prediction for object o. The training methodology for the HP network is detailed in Section 5.

The reliability of the HP network is critical to prevent unnecessary terminations of the generation and ensure that objects that would have been present are not prematurely discarded.

In all network configurations, as feature extractor from cross-attention maps, we utilize a series of four convolutional layers that bring the input from a $1 \times 64 \times 64$ to a final $128 \times 7 \times 7$ shape.

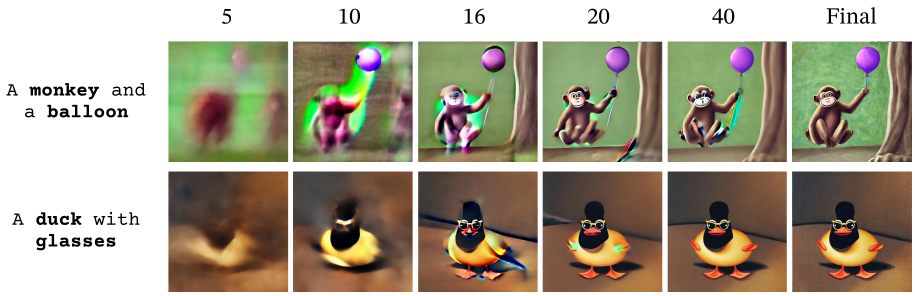

Fig. 3: Qualitative examples of the Predicted Final Image for each prompt at different critical timesteps. Already from the 16th step the final image is fully represented and the presence of objects can be predicted.

HP-R – HP-V. For scenarios where the number of critical timesteps selected is 1, *i.e.* $|\mathcal{T}| = 1$, we have explored two distinct backbone options: Resnet50 and Vision Transformer (ViT), which correspond to the HP-R and HP-V architectures respectively. These architectures are tailored to process the PFI to extract features that are then combined with those extracted from the cross-attention maps to get the final prediction. HP-R concatenates the output of the Resnet backbone on the channel level with the processed cross-attention maps and a further two-layer CNN is used to merge channels down to 512. HP-V processing outputs a vector that is concatenated to the flattened version of the cross-attention maps output. Both networks employ a final linear layer for the binary prediction.

HP-MultiR. For scenarios where $|\mathcal{T}| > 1$, we have created the HP-MultiR network in which Resnet50 was used to extract features from different timesteps in parallel. Features and the processed cross-attention maps are concatenated on the sequence dimension before applying two Conv3d layers with a final 3D pooling to reduce dimensions. A final classifier is eventually used for the final prediction.

HP-A. Additionally, we developed the HP-A network to specifically investigate the influence of attention maps on hallucination prediction. This network configuration excludes the PFI from its input, focusing solely on the features extracted from attention maps. A final prediction layer is attached directly to the common cross-attention maps feature extractor. By employing a similar architecture with convolutional layers as the other HP variants, the HP-A network focuses on evaluating how well attention maps alone can predict hallucinations. The results from this model provide critical insights into how effectively attention maps alone can inform the hallucination prediction process in diffusion models.

5 Hallucination Network Training

Dataset Creation. To train the Hallucination Prediction network we collected 900 prompts obtained by combining 75 distinct animal subjects with 12 objects

with the prompt "`A {animal} and a {object}`". To augment the dimensionality of the dataset and thoroughly investigate output variations influenced by different seeds, we generated 12 images for each prompt using distinct seeds. Following this protocol we generated nearly 10.000 images by making use of Stable Diffusion v2.0 generator [24]. During generation, we fixed 50 steps of the diffusion process and we collected the PFI and cross-attention maps A at multiple time steps[4].

Target Objects Extraction. While our dataset comprises prompts with objects in predetermined positions for simplicity, we integrated object extraction to simulate real-world scenarios. For this purpose, we employed gpt-3.5-turbo-1106 [18], selected for its robust zero-shot generalization abilities. This method stands in contrast to conventional text tagging techniques that generally necessitate specific training for each domain.

The extraction procedure is time-efficient and can be executed concurrently with the initial diffusion steps. Details on the specific prompts used in this study can be found in the supplementary materials.

Label Creation. An essential feature is the development of an automatic labeling system to confirm the presence of particular objects in the generated images. This system must function without a fixed set of object labels, requiring the adoption of an open vocabulary approach. To achieve this, we adopted OWLv2 [17], an open vocabulary detector renowned for its robust detection capabilities and for providing confidence scores for each identified object.

6 Time Saving Analysis

Our study primarily explores the time-saving benefits of the HEaD approach in DMs when trying to generate a *complete* image. In our analysis, we found that accurate generation of both objects in complex scenarios was achieved in only 59% of cases by SD2 without HEaD. This statistic underscores the challenges models encounter when generating multiple objects accurately, particularly as the complexity of the prompt and object combinations increase. Certainly, with more objects and increasingly complex prompts, the probability of correct generation diminishes, which in turn heightens the impact of HEaD on time-saving.

In the dataset, which is made with 12 seeds per prompt, we found that every prompt successfully led to at least one correctly generated image, and 98.4% resulted in at least three accurate images. This confirms the feasibility of the prompts for some random seeds. HEaD serves here as an implicit evaluator, swiftly identifying instances where the generated image is likely to be inaccurate. By promptly halting these less promising generative paths, HEaD allows for more efficient use of resources, enabling quicker initiation of new generation attempts with different seeds.

[4] In particular, the critical steps $\mathcal{T}$ are chosen as follows: [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 14, 16, 18, 20, 25, 40]

6.1 HEaD impact on Time Saving

HP Performance. Labels in our dataset are created using an open vocabulary detector, which assesses whether each object is present (1) or absent (0) in the images. The HP network, based on these labels, decides whether to continue or halt the image generation process. When a True Positive (TP) occurs, the correct generation proceeds uninterrupted, having no effect on computation time. Conversely, a False Positive (FP) allows an incorrect generation to continue without interruption, thus missing an opportunity for time savings, but still not impacting computation time. A True Negative (TN) indicates an incorrect generation has been correctly halted, leading to time savings. Finally, a False Negative (FN) means a correct generation is mistakenly stopped, resulting in a loss of time.

Thus, in order to save computational time, the network should be trained to balance both high recall and a high TN-rate. High recall ensures the HP network effectively identifies all instances of correct generation, minimizing FNs and avoiding unnecessary termination of accurate processes. Simultaneously, a high TN-rate boosts the HP network's capability to maximize true negative outcomes, allowing for early termination of incorrect generations by accurately identifying cases where not all requested objects in the prompt are included. This dual focus on both recall and TN-rate optimizes the generation process by reducing time loss, yet still maintaining the quality of the output.

Critical Timesteps selection. The ratio $\frac{t_{c_k}}{T}$, where t_{c_k} denotes the latest t in the set of critical timesteps $\mathcal{T}$ and T is the total number of steps in the generation process (with 50 being the standard for SD), plays a crucial role in determining the percentage of time saved. An earlier detection, indicated by a smaller t_{c_k}, can potentially lead to greater time savings in case of a correct hallucination identification. However, this scenario presents a significant challenge: in the initial stages the quality of attention maps and PFIs is lower. This lower quality affects the performance of the HP network resulting in reduced recall and TN-rate, as shown in the plot in Fig. 4. Therefore, this tradeoff between early detection and maintaining the quality of attention maps and PFIs is essential for maximizing the efficiency of the HEaD approach.

Finally, to quantify the time saved or lost using the model, we conducted Monte Carlo simulations based on the models presented in the next section. The algorithm calculates a savings of $\frac{t_{c_k}}{T}$ of the generation time when a true negative (TN) is detected. Conversely, it accounts for a time loss when a new restart is necessary due to a false negative (FN). The detailed algorithm and simulation results are provided in the supplementary materials.

7 Experimental Results

The evaluation of various HP Network variants underscores their influence on the image generation process. The computed Recall and TN-rate metrics, which are influenced by the t_{c_k} value, serve as key indicators of model performance. As

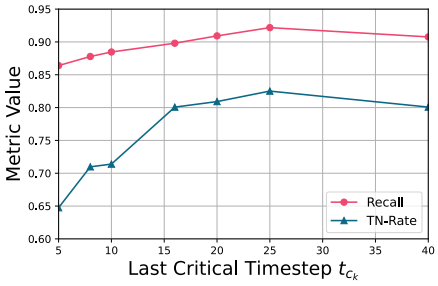

Fig. 4: Recall and TN-rate values for HP-R across various t_{c_k}. Lower t_{c_k} values, associated with lower quality input, significantly impact the TN-Rate but minimally affect Recall. Consequently, the overall time saved tends to be greater for smaller t_{c_k} values.

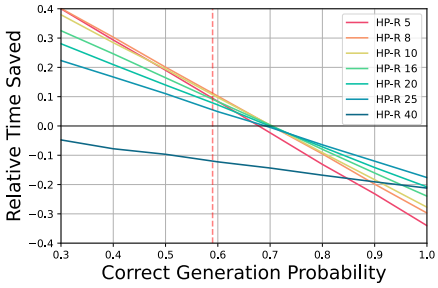

Fig. 5: Relative time saving between adopting or not the HEaD approach to reach a *complete* generation, using HP-R with different t_{c_k}, depending on the probability of a correct image generation. The vertical red line marks the probability of correct generation in a two-objects scenario, *i.e.* 59%.

depicted in Fig. 4, the TN Rate typically increases with a higher t_{c_k}, whereas Recall tends to remain stable across different stages of detection.

To provide a final efficiency assessment, the percentage of time saved during generation has been adopted as the primary metric for final comparison. This metric integrates the effects of varying t_{c_k}, Recall, and TN-Rate values, offering a quantifiable measure of each model's effectiveness in reducing generation times. For these experiments, a correct generation probability of 59%, as derived from the dataset, has been employed to ensure accuracy in the evaluations. Table 1 provides a comparative analysis of HP-R and HP-V, illustrating the time saved when these networks operate at different t_{c_k} intervals. Both networks have the highest impact when $t_{c_k} = 8$, where HP-V saves up to 12.66% of generation time. Higher t_{c_k} values can enhance input quality and metric results, but they may limit time-saving opportunities. No models bring any benefit when using $t_{c_k} \geq 25$, as the time saved in case of a correct prediction is insufficient.

In Fig. 5, an analysis is presented to illustrate the relationship between the relative time saved and the generation probability across different t_{c_k} values. The vertical line indicates a 59% correct generation probability, typical for scenarios involving two objects, as observed in our dataset. More complex prompts, which often require synthesizing additional objects, tend to have lower probabilities of achieving a *complete* generation, thus enhancing potential time savings. Notably, $t_{c_k} = 8$ offers the optimal balance, providing significant time savings, especially when the probability of *complete* generation is as low as 40%, where time savings can reach up to 30%. Conversely, when the probability of a *complete* generation is high, using $t_{c_k} = 5$ results in considerable time loss due to imperfect Recall, which can prematurely halt a correct generation. Additionally, employing HEaD at $t_{c_k} = 40$ provides no benefits in any scenario, as the time saved in the rare

$\mathcal{T}$	**HP-V**	**HP-R**
5	9.7	9.11
8	**12.66**	**10.56**
10	9.68	10.34
16	6.72	8.93
18	5.77	5.78
20	5.75	7.25
25	-0.35	5.32
40	-14.11	-11.67

Table 1: Percentage of time saved for all models. t_{c_k} is the last diffusion timestamp considered over the 50 of SD2.

Model	$\mathcal{T}$	**% Time Saved**
HP-A	10	6.65%
	16	3.04%
	20	-0.73%
HP-Multi	6-8-10	-3.72%
	10-12-14	8.99%
	16-18-20	6.88%

Table 2: Percentage of time saved for HP-A and HP-Multi in different $\mathcal{T}$ scenarios.

event of a true negative is merely 20%, considering the 50 steps generation pipeline of SD2.

In Table 2, HP-A testing serves as an ablation study to underscore the significance of Predicted Final Images. In the absence of PFIs, which are unique per image and not per object, the HP-A model shows a marked decrease in its ability to detect early hallucinations and thus in time saved. With $t_{c_k} = 10$ only 6.65% of generation time is saved.

The HP-Multi model takes an advanced approach by focusing on multiple $\mathcal{T}$. A noteworthy aspect of HP-MultiR performance is its effectiveness in later timesteps ($t_{c_k} = 14$), compared to a less marked performance in early timesteps. This discrepancy can be attributed to the inhomogeneity of the data in the early stages, where the characteristics of the data change considerably from one step to the next. This variability makes the mixing of the features in these early stages less effective. In contrast, data in later stages tend to be more uniform and stable, allowing for more effective learning and integration of features from multiple time steps, thus improving model performance.

8 Conclusions

This paper introduces HEaD, an innovative approach that not only enhances the efficiency and accuracy of image generation with Diffusion Models but also significantly reduces computational resources. A key innovation is the Predicted Final Image, an effective early error prediction indicator when used in conjunction with cross-attention maps. The effectiveness of our framework in saving time is closely tied to the recall and TN-rate of the Hallucination Prediction network, highlighting HEaD's capacity to improve image generation in a variety of complex scenarios.

HEaD represents a preliminary step in exploring the sustainability and effectiveness of diffusion models, especially for large, complex datasets. Looking

ahead, we are committed to further advancing this field of study also by collecting larger datasets with more target objects and more complex visual prompts and proposing challenges for the scientific community to test better early detectors.

9 Acknowledgment

This work was supported by the MUR PNRR project FAIR - Future AI Research (PE00000013) funded by the NextGenerationEU, the PRIN project CREATIVE (Prot. 2020ZSL9F9), the EU Horizon projects "European Lighthouse on Safe and Secure AI (ELSA)" (HORIZON-CL4-2021-HUMAN-01-03), co-funded by the European Union (GA 101070617) and "ELIAS - European Lighthouse of AI for Sustainability" (No. 101120237). Further, we thank G. Fiameni (NVIDIA) for helping with the generation of the dataset.

References

1. Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Zhang, Q., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., Karras, T., Liu, M.Y.: eDiff-I: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers. ArXiv (2022) 4
2. Betti, F., Staiano, J., Baraldi, L., Baraldi, L., Cucchiara, R., Sebe, N.: Let's vice! mimicking human cognitive behavior in image generation evaluation. In: ACM Multimedia (2023) 2, 3
3. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models. ACM Transactions on Graphics (TOG) (2023) 1, 2, 4
4. Dhariwal, P., Nichol, A.: Diffusion Models Beat GANs on Image Synthesis. In: NeurIPS (2021) 1
5. Feng, W., He, X., Fu, T.J., Jampani, V., Akula, A.R., Narayana, P., Basu, S., Wang, X.E., Wang, W.Y.: Training-Free Structured Diffusion Guidance for Compositional Text-to-Image Synthesis. In: ICLR (2023) 2
6. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Networks. NeurIPS (2014) 1
7. Helbling, A., Montoya, E., Chau, D.H.: ObjectComposer: Consistent Generation of Multiple Objects Without Fine-tuning. arXiv preprint arXiv:2310.06968 (2023) 2
8. Hertz, A., Mokady, R., Tenenbaum, J.M., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-Prompt Image Editing with Cross Attention Control. arXiv preprint arXiv:2208.01626 (2022) 4
9. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In: EMNLP (2021) 3
10. Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models. In: NeurIPS (2020) 5
11. Hu, Y., Liu, B., Kasai, J., Wang, Y., Ostendorf, M., Krishna, R., Smith, N.A.: Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20406–20417 (2023) 3

12. Karthik, S., Roth, K., Mancini, M., Akata, Z.: If at First You Don't Succeed, Try, Try Again: Faithful Diffusion-based Text-to-Image Generation by Selection. arXiv preprint arXiv:2305.13308 (2023) 2, 4
13. Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes. In: 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings (2014) 5
14. Liu, N., Li, S., Du, Y., Torralba, A., Tenenbaum, J.B.: Compositional Visual Generation with Composable Diffusion Models. In: ECCV (2022) 2
15. Lu, Y., Yang, X., Li, X., Wang, X.E., Wang, W.Y.: LLMScore: Unveiling the Power of Large Language Models in Text-to-Image Synthesis Evaluation. In: NeurIPS (2024) 2, 3
16. Mao, J., Wang, X., Aizawa, K.: Guided image synthesis via initial image editing in diffusion model. In: ACM Multimedia (2023) 4
17. Minderer, M., Gritsenko, A., Houlsby, N.: Scaling Open-Vocabulary Object Detection. In: NeurIPS (2024) 9
18. OpenAI: Gpt-4 technical report. ArXiv **abs/2303.08774** (2023) 9
19. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023) 1
20. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision. In: ICML (2021) 5
21. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv preprint arXiv:2204.06125 (2022) 1
22. Rassin, R., Hirsch, E., Glickman, D., Ravfogel, S., Goldberg, Y., Chechik, G.: Linguistic Binding in Diffusion Models: Enhancing Attribute Correspondence through Attention Map Alignment. In: NeurIPS (2024) 4
23. Reed, S.E., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: ICML (2016) 1
24. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-Resolution Image Synthesis with Latent Diffusion Models. In: CVPR (2022) 1, 4, 5, 9
25. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015) 5
26. Samuel, D., Ben-Ari, R., Darshan, N., Maron, H., Chechik, G.: Norm-guided latent space exploration for text-to-image generation. In: NeurIPS (2024) 2, 4
27. Samuel, D., Ben-Ari, R., Raviv, S., Darshan, N., Chechik, G.: Generating images of rare concepts using pre-trained diffusion models. In: AAAI (2023) 1, 4
28. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In: ICML (2015) 5
29. Wang, W., Bao, J., Zhou, W., Chen, D., Chen, D., Yuan, L., Li, H.: Semantic Image Synthesis via Diffusion Models. arXiv preprint arXiv:2207.00050 (2022) 4
30. Wu, Q., Liu, Y., Zhao, H., Bui, T., Lin, Z., Zhang, Y., Chang, S.: Harnessing the Spatial-Temporal Attention of Diffusion Models for High-Fidelity Text-to-Image Synthesis. In: ICCV (2023) 2
31. Yarom, M., Bitton, Y., Changpinyo, S., Aharoni, R., Herzig, J., Lang, O., Ofek, E., Szpektor, I.: What you see is what you read? improving text-image alignment evaluation. NeurIPS (2024) 3

32. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N.: Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: ICCV (2017) 1
33. Zhang, Y., Kang, B., Hooi, B., Yan, S., Feng, J.: Deep Long-Tailed Learning: A Survey. IEEE Trans. PAMI (2023) 1, 4