One Picture and One Thousand Words: Toward integrated multimodal generative models

Roberto Zamparelli*
Università di Trento

Thanks to independent advances in language and image generation, we could soon be in the position to have systems that communicate with us by combining language and images in their output, *a skill that humans do not possess (we receive, but we do not produce images at high speed). This paper explores some of the implications of this idea: which kinds of data sets need to be developed to train such systems, in which cases language and images could be most usefully integrated and which issues could arise on the image generation and language+images integration side. Story and dialogue illustration could be relatively low-hanging fruits for this technology, and a looped combination of I2T LLMs and T2I diffusion models is likely to play a role in solving some of the issues that arise in the design of such systems.*

1. Introduction

Human beings acquire information about their environment in two ways: by witnessing it through the senses, directly or indirectly (e.g. seeing a car crash, hearing the noise, smelling gasoline etc.) or by obtaining reports through language (hearing *that* a car has crashed, *when/how* it happened, etc.). Our input may be multimodal, but the main distinction is between symbol-mediated information (language and other communication codes) and any type of (direct) sensory information.

When we turn to the *output*, however, to the way our species produces information for others to understand, the choice is remarkably narrower. We have evolved to produce language and gestures, interactively and at high speed. We have *not* evolved to be able to accompany language with images or sounds: showing the exact way in which a car crashed, or imitating the precise noise its breaks made. If we use gestures to try to convey an iconic aspect, the result is understood to be an approximation: the way I move my hand may be indicative of the way the car slid, but a million other aspects of the event reported remain unexpressed — many to be sure irrelevant, others possibly important.

Time and technology are helping us narrow down the gap between our inputs and our outputs. With enough time (and skills that many people with a good use of language often do not have) we can draw what happened, make graphs of the car's deceleration or create physical models of the objects involved. Affordable cell phone technology now allows witnesses to take pictures or videos of events, vastly expending our possibility of *showing*, not just *telling*, what took place. This, however, does not extend to what *could have* happened, or what has not happened *yet*. I can explain how (maybe, eventually)

* Center for Mind/Brain Sciences - E-mail: roberto.zamparelli@unitn.it

© 2024 Associazione Italiana di Linguistica Computazionale

Starships will land on Mars, but showing this graphically requires professional video makers and special effects.

The goal of this article[1] is to point out that AI systems from the near future will *not* have these limitations: they will be able to produce language *and* still or moving images (further into the future, 'stage noises', perhaps smells and flavors), mixing them in any way we can imagine and in many we cannot yet. This process could be fast enough to take place in *dialogue*: a chatbot could answer my questions with an image accompanied by verbal explanations, adjust the image if it is unclear, answer my questions by expanding details, link referring expressions to objects in the picture, render focus by blurring/unblurring parts, making sections transparent or adjusting colors, etc. None of these capabilities exist in humans: gestures are too schematic, photos are restricted to showing only what exists, drawing is too slow to be usable without interrupting the flow of the conversation. In short, combining language and images in on-line generation is a task in which AI systems not from the realm of science-fiction (due to independent advances in language generation and image generation, see Sec. 3) could very soon achieve superhuman abilities and produce actually useful output.[2]

This unique status comes with a down side. The aspects where NLP or artificial vision have historically progressed the fastest are those where they could leverage huge datasets, annotated via crowdsourcing & bootstrapping. Humans do create combinations of images and texts in various contexts and with various purposes, but these contexts and purposes appear to be a subset of the cases where it would be useful for an AI to generate what I am going to call I+L (Images+Language). In particular, none of these contexts and cases is at the same time fast, detailed and interactive. It follows that, before actual systems can actually be created, we need to turn our attention to three broad aspects:

- *I+L Applicability*: which types of contents would be most suited to be rendered as I+L? Which communicative goals would profit the most? What is the right I/L ratio for each type?
- *I+L Training*: What can be done to train an I+L system? Can current datasets be used, and if not, how can they be augmented?
- *I+L Interface*: How is the interface between language and images going to work? Specifically, how does reference-fixing works? How can a generated image make clear which elements come from the content specifications and which have been added to fill the scene? Can the graphic side be pressed into service to give discourse-level information?

This paper offers some preliminary considerations on these issues, but no quantitative results. The discussion here comes from the vantage point of linguistics (and image generation, to some degree). Undoubtedly, additional issues will emerge when the topic is considered as an interface design problem (human-machine interface, but also machine-machine).

1 This paper is revised and expanded version of (Zamparelli 2023)

2 Interestingly, a system that is seen answering questions by producing intertwined text and images could not be mistaken for a human production simply because no human could do it. There cannot be deepfakes here, though of course the final output could raise the same issues or authorship and attribution as other artificial creations, and give raise to the security concerns raised by its individual constituents.

The paper is organized as follows. In Sec. 2 I consider the usefulness and feasibility of an I+L system in the current NLP landscape. Sec. 3 reviews some of the AI models the I+L idea builds on, and discusses the (lack of) current datasets; Sec. 4 discusses possible situations, genres and choices that affect the creation of I+L systems. Section 5 tests current tools on a concrete example of text illustration, while Sec. 6 discusses the more difficult case of dialogue, where the image generation system has only partial information available. Some even more speculative remarks on the possible role of I+L in bridging the gap between still images and videos can be found in Sec. 7. Sec. 8 concludes.

2. Domain of applicability

The first question, as with any unproven technology, is whether an I+L generation system could be useful, and where. Specifically, which textual genres or communication types could profit the most from the integration of language and vision, and which ones should be best left to words alone?

The answer hinges on the relation between text and (related) images, which has been a topic of study in the field of semiotics (see (Martinec and Salway 2005), henceforth M&S, who build on Halliday's functional grammar and on (Barthes 1964); see (Spanjers 2021) for a more recent overview). One fundamental observation of this work is that combinations of text and images vary along a dimension which I am going to call *aboutness* (M&S's calls it, perhaps less transparently, 'status'). In an illustrated Pinocchio book, the images are *about* the text; removing them will not harm its consistency; on the other hand, removing the text is likely to hamper our understanding of the temporal and logical relation among the illustrations. At the opposite hand, a *caption* is text *about* an image (specifically, its interpretation); remove it, and the caption won't make sense. In M&S's classification, text and images can also differ in *generality* (equal, or with one being more general), in the *size* of the connecting element (the image could be connected to a single clause or to a whole paragraph; only one part of the image could be connected to the text) and in the logico-semantic relation they express (Halliday's notions of *elaboration*, *extention* and *enhancement*, see M&S pg. 349). The last relation, enhancement, corresponds to what other theories would call a meta-level of description; text passages in an art photo book, for instance, are 'about' the images in a very different way than captions: they will typically tell us how and when the images were taken, what is their significance, intention or style. An objective classification of the other logico-semantic relations seems harder to achieve, but also less important for the present purpose.

Returning to aboutness/status, M&S regard cases where the image are about the text 'as a whole' as examples of text-image *independence*,[3] and cases where each modality is about the other as examples of *complementarity*. I can think of very few examples of cases where the aboutness direction alternates (i.e. text-image combinations where, at different points, each one is about the other) — possibly due to the fact that this might confused the user. Looking at other cases where modalities are combined, we see that 'being about something' is not a necessity: in a movie, the soundtrack is not about the film plot or vice-versa, but music can be used to convey emotions or expectations. In the I+L domain this could be paralleled by the use of abstract images as illustrations (but

3 One example might be scientific posters, where space limitations force text and images to convey a common message without directly referencing each other. Obviously, complete independence would only be achieved if the image-text pairing was random. This would not result in any positive communicative effect.

also, at the typographic level, by the use of text layout, font choice, color, etc. to convey different aspects of meaning).

The starting point for automatic I+L generation is that these abstract relations are embodied differently in different existing genres, and generating the wrong text-image relation can be perceived as a violation, and probably a disturbance. Consider narrative, an obvious candidate for the I+L process. As discussed in the context of movies-from-books, depicting narrative can actually stifle mental imagery, standardize places and character, and ultimately reduce the appreciation of the story (Leitch 2007). At the same time, illustrating written narrative with still pictures is an age-old human activity; modern humans also make use of moving pictures (e.g. films), probably trading the freedom of personal mental imagery for the comfort of professional scene-setting. The duality *text* (visually free) vs. *video* (visually locked) has in fact already been shattered: virtual reality makes it possible to experience the same theatrical experience from different viewpoints (see (Palma, Schofield, and Moore 2023), on the multiple levels of viewer participation, and the Theater Elsewhere [4] app for Meta VR for an example); adventure videogames offer multiple ways of traversing a (branching) narrative, and the possibility for players to choose the likeness and equipment of their avatars. In the world of interface designs, the same tendency toward customization translates into the choice of "themes".[5] Image-augmented narrative would thus simply continue this trend, with the added value of offering the user more stylistic choice.

Storytelling is of course only one of the tasks that generation systems could tackle, and its usefulness should be evaluated in the broader context of the usefulness or desirability of AI-augmented narrative creation (Ma, Qiao, and Liu 2024). More practically-minded targets for I+L systems could be how-to instructional videos (e.g. "*How to replace the battery of your phone.*", "*How to make a strudel.*"), descriptions of longer activities (e.g. a mountain walk created by combining 3D map imagery with a commentary on the most difficult steps; a road trip guide that shows possible stops and their attractions; an interactive surgical training session); renderings of complex objects and their operations (e.g. the 3D view of a device extracted from a patent description; a CAD mechanical model that offers the possibility of discussing, interactively, the function of each of its subparts, shown from different angles and in action; the mechanics of a virtual board game). Finally, they could become a tool in the development of multimodal AI: for instance, a language+vision model might require I+L to explain its own output in Chain-of-Thought fashion.

Most of these use cases highlight the importance of a *dialogic* dimension, compared to a static I+L system. One neighboring domain which has already received considerable attention is the interaction needed for *image refinement*. Current text-to-image models create images which may be far from their creator's expectations. High-level linguistic instructions (e.g., in Fig. 3: "*Make the president smaller and move it closer to the steps.*") would be a natural way to refine the result. This could of course be a Text2Image-only task (T2I: the system answers by modifying the image, see TiGAN, (Zhou et al. 2022)), but since image creation is (and is probably going to remain) computationally more demanding than language generation, it is easy to see the usefulness of verbal interaction, in clarifications ("*Should the seat be smaller, too?*") or explanations (User: "*What are the bright things on the president's chest?*"; System: "*Watch chains.*"; "*Too many,*

4 `https://shorturl.at/tDRT5`

5 In these last cases, the choice is typically offered by presenting compositional alternatives ("***THIS** style of hair or **THIS** style?*"); one issue to investigate is how a menu-style choice among different visual versions could be augmented, or replaced, by a textual dialogue that converges onto a desired style.

remove all but one.", "*Which one?*"). Note that in this case, unlike in the narrative example above, language is not a part of the final product (an image, or a video) — just a tool at the service of a more effective image creation process.

3. Background: I+L precursors and training

The possibility of generating I+L builds on recent advances in a number of neighbouring tasks: language generation (OpenAI's GPT family (Brown et al. 2020), Anthropic's Claude family, Google's Gemini family, Meta's Llama family), text-to-image (T2I) models (GLIDE, (Nichol et al. 2022), DALL-E (Ramesh et al. 2022), Stable Diffusion, etc.), text+image-to-image (T+I2I) generation (Lu et al. 2021), T2I and I2T with the same model (Yu et al. 2023) and multiple-round image editing with language (Zhou et al. 2022). Many of these advances have been made possible by a move from annotated images datasets such as ImageNet (Deng et al. 2009) or MS-COCO (Lin et al. 2014) to image embeddings like CLIP (Radford et al. 2021), trained on image-text matching collected on a huge scale. CLIP does not have a limited vocabulary and has been shown to improve VQA tasks (Shen et al. 2022), including image captioning (Ghandi, Pourreza, and Mahyar 2023), another task quite relevant for I+L.

Other advances came from the idea of applying to image generation the textual embeddings developed by recent, large-scale language models (e.g. T5 in IMAGEN, (Saharia et al. 2022)), combined with image upscaling. Given the strict connection between language and images required by an I+L system, another relevant strand of research is the one that aims to go from images to descriptions longer than typical user-generated captions, either by associating captions to the individual bounding boxes (*Dense Captioning*, (Johnson, Karpathy, and Fei-Fei 2016)) or by creating extensive human-readable descriptions from images (*Image Paragraph*, (Krause et al. 2017)). The aim of some of this work is assistive technology for the visually impaired (Fernandes et al. 2022), but they can be useful to map linguistic descriptions of entities back to the images, or as part of a circular flow that goes from language to images and back to language (see Fig. 8 below).

In the last two years the field has seen an extremely fast growth, with new open-source data made available monthly (e.g. the CLIP-based LAION-5B dataset), ongoing efforts to make models trainable at lower costs (Li et al. 2023; Evans et al. 2024), and the release of open-source T2I software for low-end hardware (Stable Diffusion[6]), which is creating a huge pool of testers for future models. A recent survey on systems that combine text and images in some way or other (He et al. 2024) lists almost 400 relevant papers just in the last 2 years, distinguishing four main use categories: generating or editing a combination of image and text from text and sometimes images (see e.g. (Dong et al. 2024)), using language to organize the layout of visual elements, refining an image's linguistic prompt, and evaluating the quality of an image. In the vast majority of the cases considered, the goal is to produce *images* (from text or text+images), or *text* which either describes the image (automatic captioning) or uses the image as ground truth (as in in VQA, (Antol et al. 2015)) — all tasks that can be generalized from innumerable examples of text-image associations: human-generated captions, textual contexts and image metadata.

The goal of the I+L system envisioned here, however, is different: the visual and the linguistic side of the output should provide contents that can span the full range of

6 https://stability.ai/blog/stable-diffusion-public-release

aboutness relations discussed in Sec. 2, including the possibility of switching from text-about-the-image to the opposite, and the case of complementarity, where the images are connected to the text as a whole, and the two sides provide different, jointly relevant information. There have been attempts at creating datasets where the text and images are complementary (see BD2BB, (Pezzelle et al. 2020)), but there are many more ways to be complementary than to coincide, and systematic ways to generate training materials specific for I+L has yet to be found. At the time of writing, the *Tasks* sections of two of the largest dataset repositories (Paperswithcode Datasets[7] and Hugging Face[8]) do not contain any task labeled "image *and* text generation" (though Huggingface has a generic "any-to-any" category). The OpenLEAF project (An et al. 2024) explicitly aims at creating a benchmark dataset of combined text and images, but this is generated by asking GPT4 to create a textual narrative interspersed with image descriptions, which are then turned into full T2I prompts and fed to StableDiffusion (SDXL); the result is evaluated by BingChat on ground of image relevance and consistency. The resulting dataset contains data from genres such as visual instruction generation, story generation and rewriting, automatic generation of web pages and posters; most of these tasks start with *complete, written* texts; none seems to involve dialogue, where old images can conflict with new information coming in (see Sec. 4.4). GPT4 has never seen the type of mixed, interactive I+L uses that AI could make possible, so it can hardly be expected to be particularly innovative on the generation side.

One domain where humans routinely combine oral production and images is presentations with slides. Large public slide repositories do exist (e.g. Slideshare-1M[9]), but no dataset that combines the speaker's voice (or its transcription) with the slide presented. TED Talks, another possible model of a certain way of combining speech and images, have been compiled in various machine-friendly forms, but not in one that systematically associates the slides used with the speech (TED Talks[10] has voice and videos of the speaker's body; TED LIUM 3[11] only audio and transcripts).

Another potential source of training material are instructional videos (on products, procedures, DIY projects, etc.), which could be useful as long as they do no not (mostly) show the speaker. Turning them into I+L-training material would involve reducing the video to maximally informative still frames of the type/resolution that could be matched by current T2I systems, then linking them to the corresponding voice passages. Multiple video frames could be used to generate a 3D representation of the objects shown (e.g. using NeRF-like techniques, (Mildenhall et al. 2021)). In this case, the resulting dataset would contain a paired combination of text/voice and the 3D embeddings of the entities mentioned at every point. An I+L system could learn to accompany the text passage with individual 3D images, giving the listener the possibility to explore them interactively (asking for rotation, zooming, etc.). Its usefulness would crucially depend on the accuracy with which the entities described can be mapped onto existing 3D object datasets (e.g. Objaverse[12]).

Next in the list of possibilities we have commercial *movies*, and *comics*. When the former are combined with storyline datasets like MPST (Kar et al. 2018), it may be possible to extract still frames that best highlight specific storyline passages to be used

7 https://paperswithcode.com/datasets
8 https://huggingface.co/models
9 https://purl.stanford.edu/mv327tb8364
10 https://paperswithcode.com/dataset/ted-talks
11 https://paperswithcode.com/dataset/ted-lium-3
12 https://github.com/allenai/objaverse-xl

for I+L training. Unlike with videos, it remains to be seen if a movie's storyline is not too coarse to be meaningfully mapped to the still frames. Comics are in a way the literary genre that comes closest to the possible output of a type of I+L system, and there have been studies on how the structure of comics narrative mimics language (Cohn 2013), down to many similarities at the cognitive level (Lindfors et al. 2024). The problem, in this cases, is in the limited range of topics and themes that the comics literature covers (especially if one wants to exclude copyrighted material). Comics could be useful to train an I+L system that produces just that — comics. Their utility at large remains to be proved.

Finally, there is ongoing work on the one type of visual output that humans do produce at speed: gestures. The practical goal, in this case, is to produce natural gestures for embodied conversational agents (avatars and humanoid robots) (Liu et al. 2021). In this case, ML models can be trained on actual human data, even in the absence of a full understanding of the semantics of gestures. The results of a recent challenge (Kucherenko et al. 2024) shows that current models can produce gestures which humans judge natural, but often not appropriate for the specific discourse context. Moreover, there is still no agreement on how the evaluation should be carried out (Wolfert, Robinson, and Belpaeme 2022), partly due to the fact that co-speech gestures fall in four classes: beat, deictic, metaphorical and iconic (McNeill 1992), which may be hard to disentangle. The latter two classes (e.g. using open hands to signify emptiness, or distancing the hands to indicate the size of an object) are those that come closest to the use of generated images that I am envisioning in this paper, and could offer important cues on the placement of images in a full I+L computational system.

4. Parameters in I+L creation

Creating new datasets or helping the systems to best use the existing ones requires a detailed theory of how the images and language should interact. Assuming we start from a text without images, (at least) four aspects decide the actual generation process.

(1) a. **Desired Text/Image Ratio**
How much of the text should be turned into images? Should the images *accompany* the text, or *replace* it?

b. **Mined vs. Generated Images**
How many of the images to be added are preexisting images (of actual objects or people) and how many should be generated in addition? (from the text alone, or from the text plus the preexisting images)

c. **Exhaustive / Partial Text Knowledge**
Does the system have full knowledge of the text, its purposes and its composition details, so that images can be planned as a coherent body, or is some of this information missing?

d. **Consistency Level**
What is the tolerance of the generated images, i.e. how inconsistent in details, style, or viewpoint can they be?

I expand these points in the following subsections.

4.1 Text/Image Ratio

The ideal images/language ratio depends on the genre and should probably be a user-specified feature. Human (off line) examples go from illustrated novels (text, for the

most part) to the intertitles that occasionally interrupt silent movies. One big advantage of a generative I+L system over predefined text and image combinations is that the I/L ratio could be changed on the fly: the images could be tuned up or down depending on how useful their contribution is felt to be. On the other hand, if the starting point is text, an automatic system might tend to concentrate its images on those parts which have a higher degree of *imageability* (a measure related but not identical to *concreteness*, see (Della Rosa et al. 2010; Ljubešić, Fišer, and Peti-Stantić 2018)), which might however concentrate in specific subsections of the text, creating an undesirably 'lumpy' distribution. Using imageability as the main cue risks focusing on secondary parts of the plot line (in fiction), or on irrelevant details (in instructional videos, imagine been treated with photos of the company logo or headquarters in a discussion of how to repair one of their products).

Another fundamental decision is whether images should *supplement* the text, as in book illustrations, or *replace* it where it becomes redundant. As artificial video generation (e.g. OpenAI's SORA[13] or Runaways Gen-3[14]) becomes mainstream, one can imagine the creation of longer videos as a process in which textual sequences are progressively replaced by visual material (plus words, when needed), but it is also possible to think in reverse and use language to add thoughts or comments related to an image. Current captioning systems already volunteer comments that go beyond the content (e.g. remarks on an image's style or atmosphere). Some of these comments could find their place back in the image, as annotations.

A related issue is whether the text that accompanies the images should be written within the image (in the form of captions, or comics balloons), appear before or after the images (like intertitles in silent movies) or be voiced as if by a narrator.[15] The advantage of ballons over captions is that they clarify who is talking. In films, this is achieved using the voice timber of different actors, and close ups (see (Arijon 1976, Ch. 5)). In computer-generated still images, this could be achieved by highlighting one of the characters or via other graphic manipulation (see section 7).

4.2 Generated vs. Mined

The simplest way to create I+L does not involve any image creation, just image selection. One can imagine a system that selects a set of imageable nouns or noun phrases found in the input text, searches image repositories for illustrations of these words and either inserts them directly close to their mention or suggests them to the user (this already happens for text-driven emoji selection in keyboard software like Microsoft SwiftKey; the ability to select existing images matching a conversation was already demoed by Amazon in its re:MARS 2022 keynote).

Putting aside the legal concerns associated with the reuse of copyrighted materials, the selection of photographic material would alleviate the problem of stylistic consistency (photos of objects can be stylistically different, but not as different as drawings) without completely solving it. Unfiltered augmentation of a text discussing a product with photos of the product might seem reasonable, but the photos could end up showing different models, or include images of the product under repair. Similarly, adding

13 `https://sora.aitubo.ai/`
14 `https://app.runwayml.com`
15 It is worth keeping in mind that voice-over narration is not very popular is modern cinematography, possibly on the account that it breaks the narrative identification. But it is not clear that narrative identification is a desideratum in many types of I+L dialogue.

open-source photos of public characters to a text mentioning them could be a bonus, but selecting the person at different ages would be confusing.

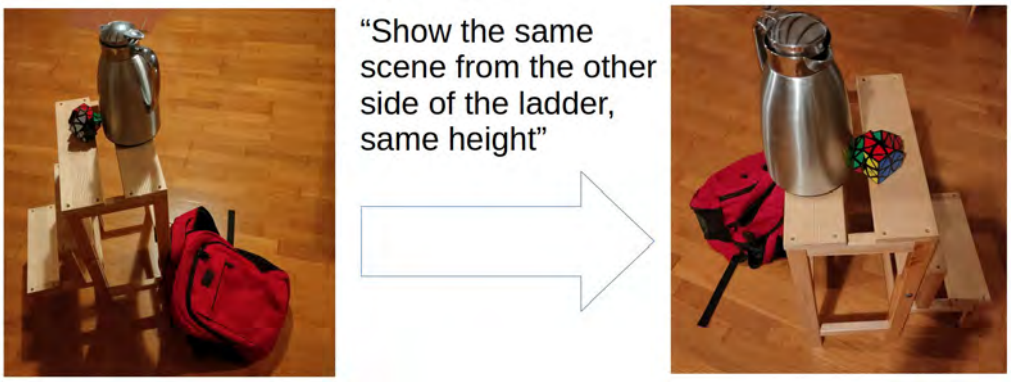

Figure 1
The problem of viewing angles: different takes of the same complex scene. Currently unsolved in a general form.

It is reasonable to assume that actual I+L systems will feature a mix of real and generated images: real images provided with the original text or retrieved from it, acting as *grounding*; generated images, to embed the objects in the context created by the narrative, or recasting it in different angles, situations and levels of detail. This can work, of course, only if a T+I2I generation system is capable of creating images which can reliably render a textual description preserving the style, characters and environment of one or more reference images. (Ruiz et al. 2023, DreamBooth) and (Gal et al. 2023) describe approaches to generate the same object in different contexts, which work by associating a small set of specific images to a unique token identifier. For instance, in (Ruiz et al. 2023), "[S]" is a unique identifier and the T2I is fine-tuned to associate "The [S] dog" to a particular dog image, while preserving the association of "the dog" to a more diverse set of dog images. While influential, this approach faces two issues: how to carry out this process starting from a single reference image, and how to adapt it to a situation where new objects constantly appear and are picked up by the narration, considering that the algorithm requires repeated fine-tuning of large T2I systems. As it stands, this conditioning seems particularly unsuitable for the on-line cases (dialogue, partial text input) described in section 6. More recent approaches (Tewel et al. 2024) avoid the fine-tuning step and focus instead on sharing relevant attention heads across images, obtaining a good combination of object consistency and prompt fidelity. Yet another approach involves treating the generation of additional images from the mined reference as a sequence of picture editing steps. Recent work on text-conditioned image editing can make objects in the image change pose or alter their style (see (Kawar et al. 2023, Imagic)), but it cannot yet show a completely different viewpoint of the same image, as in Fig. 1 (which might be a problem also for (Tewel et al. 2024)). One extreme solution might be to map the scene onto a set of objects (taken from a repository of 3D objects like Objaverse), place them in a 3D space that can be manipulated by 3D drawing engines like Blender, alter the viewpoint, then map the result back into photo-quality using methods such as (Liu, Xu, and Chen 2023).

It is important to remember, however, that movies have cuts, and that their characters and objects tend to reappear in different situations. This makes the problem of delivering object identity in changing scenarios a central issue for video generation;

Figure 2
Different realizations of "A large ceramic mug with white enamel lip, a left-protruding rounded handle and the drawing of a Japanese wave within a circle, which takes two thirds the visible side. Magazine photograph." **Top:** Stable Diffusion; **bottom:** DALL-E 3

Figure 3
Prompt: "Abraham Lincoln, seen from three quarters on a high bench, meets delegates from Virginia that bring him an appeal. Photographic color illustration, solemn. Faces well visible." DALL-E 3

as the length of generated videos increases, this issue is likely to receive a lot of well-funded attention outside of the NLP community. It would thus be wise to gratefully await for their solutions.

4.3 Consistency

Current T2I generators have a considerable margin of variation even for well-specific categories (Fig. 2), due to the underspecification of textual information and to the non-deterministic nature of the generation algorithms (e.g. diffusion models). As discussed, this creates a problem of consistency among the various images, but an even more pressing issue of consistency with the original text.

T2I generators take in input carefully engineered prompts, which are typically complex noun phrases. If the starting point of I-L generation is sentences or paragraphs extracted from the original text (say, those selected for their high imageability), the quality of the generation process can be expected to degrade. To make up for the lack of image-specific information, the generator will try to add plausible material derived from its training. Consider Fig. 3: the generator has more or less managed to render Abraham Lincoln's appearance,[16] but has made up the faces of the delegates, their number and the place where the encounter takes place. In a different terminology, Lincoln is a specific *type* (a politician, a president) and a specific *token*; the delegates are type-specific but not token-specific; the hall is neither type nor token specific. I call the latter kind of object *fillers*: elements not mentioned in the prompt but added to make the illustration 'nicer'. The problem is that these details could become inconsistent with the development of the plot, and with later illustrations based on it.

16 Not a given: in a follow up with a simpler prompt ("Abraham Lincoln meets delegates from Virginia. Photographic color illustration, detailed."), DALL-E-3 rendered Lincoln as a black man, while Stable Diffusion gave to all delegates the face of Lincoln.

A special problem is raised by hyperonymy. A story about a dog of unspecified breed should prompt the AI to draw a particular type of dog, but then the dog must then remain *of that breed* in any subsequent image. In this case the type is specific, the subtype is not, thus becoming a filler.

A strategy to cope with fillers could be starting the processing by rounding up the set of characters, objects or places that recur in the whole text ("participants"); when they are derived from reality, searching for exiting images and using those for grounding; when the participants are made up, making up corresponding images, keeping them consistent with the style, period or location of those already present. These initial image sets (real or made up) can be used to condition the remaining illustrations, with Ruiz' DreamBooth method described above or with analogous technologies.

The price to pay when the images generated are inconsistent is that the user becomes confused and the graphic side stops being an asset in the fruition of the text and becomes a disturbance. Inconsistencies are not limited to object recognizability. Unexpected visual transformations throughout the story line, such as switching to the mirror-image of a recent scene or changing graphic style can be equally confusing, if administered without a rationale. Aspects like changing angle and viewpoint are essential ingredients of the basic grammar of image sequences and films (Arijon 1976), and they are one of the creative aspects over which the user of an I+L system might want to retain control. It should also be noted, however, that *style change* is more a cultural than a cognitive source of disturbance: reality as we perceive it does not come in 'styles'. Style change from one scene to the next has started to be employed to good effect in comics and animated movies. Once again, its usefulness lies in the possibility of controlling it.

4.4 Exhaustive vs. Partial

At an epistemic level, the input that an I+L system could receive follows in two categories: *complete* and *incomplete information*. In the first, a system receives the whole text (ideally, along with information about the time and circumstances of its composition, plus of course the user's desiderata), and can analyze it before starting the generation process. In principle, this should give the opportunity of using graphic ingredients that will never be contradicted at later points. In the second case, the system does not know how the text is going to evolve, so any representation added at time t might contain aspects that are contradicted at time $t + 1$. These two cases call for different strategies:

- In the *complete* information scenario, the best approach should be to select all the sentences to be turned into images and expand them into fully detailed prompts, using information gathered from other parts of the text. This should in principle avoid inconsistencies, and it could ameliorate the problem of accidentally creating images from text portions that have structures far from the usual T2I prompts (e.g. questions, imperatives, hypotheticals, disjunctions).
- In the *incomplete* information case, two strategies seem worth pursuing, one monotonic, one not.
  - MONOTONIC: creating *minimal graphic representations*, sketching only what is explicitly mentioned in the prompt, with no fillers and maximally non-descriptive participants. Fill out details as new

information emerges (keeping as much as possible what is already there).
- NON-MONOTONIC: *thematize changes.* Creating images with the normal amount of fillers. As new information arrives that contradicts what has been drawn, change it and use language to draw the user attention on what is happening.

In the following section I will try to sketch a pipeline for the first situation (complete information) using an as example passages from the story "The Legend of Sleepy Hollow"[17], by Washington Irving. Obviously, narrative text illustration is a very low-hanging fruit for an I+L system, and one for which it is easy to imagine a reasonable gold standard: an evaluation test set could be built from illustrated books, removing the images and marking the narrative points for which they were generated; the I+L model would be requested to generate images for the same portions, and the result could be compared with the originals in terms of contents (using automatic captioning on both sets) and graphic quality (asking readers). Even this simple task, in my opinion, contains elements that could be used in more complex I+L forms. In section 6 I turn to a brief discussion of partial information cases, though I have nothing to add at the moment on the "thematize change" approach.

5. Adding images to a complete document

A schematic pipeline to automatically add images to (narrative) text is shown in Fig. 4. To test it, the original was tokenized and parsed, then annotated with imageability indexes from the MAGHAR predicted Imageability Lexicon (Ljubešić, Fišer, and Peti-Stantić 2018). The sentences were ordered by imageability/length ratio. This being an exploratory study, no attempt was made to select sentences that were sufficiently spaced in the original, or to distinguish between main and secondary clauses, or between main narrative line and subplots (probably none, in this case).

The next step was to feed sentences with high imageability index to an image generation software. I experimented with DALL-E 3, but the prompt needed to generate the portrait in Fig. 4 was

> A man in a rural part of 19th century America, described as: "His head was small, and flat at top, with huge ears, large green glassy eyes, and a long snipe nose, so that it looked like a weather-cock perched upon his spindle neck to tell which way the wind blew." Photo portrait.

Removing the first part of the prompt led to non-human caricatures. Combining the final picture with the text passage gives us a first shot at an illustrated text (Fig. 4, output ②). To push I+L further, we need to remove the text used to generate the image from the original. This could be done by keeping track of which sentences have been fed to the image generator and removing them, but this approach could easily cut anaphoric connections, decreasing the coherence of the text as a whole. I opted to pass the generated image to GPT4o and ask for a caption. The same model was then prompted to remove from the original text the sentence which matched this caption most closely, leaving a fluent text. The result (Fig. 4, output ③) seems convincing, if

17 Obtained from `https://www.gutenberg.org/files/41/41-h/41-h.htm`

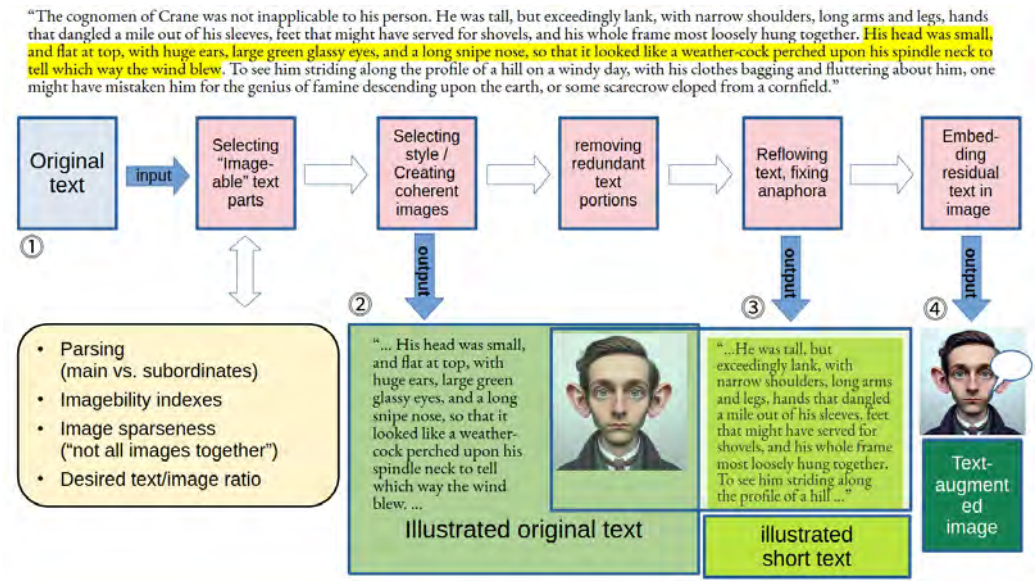

Figure 4
Going from an input text to different I+L output scenarios.

unproblematic. The LLM was also found to be capable of pruning the original sentence, removing only the descriptive material mentioned in the caption and leaving the rest as a readable passage.

To explore the limits of automating the process just described I used a combination of GPT4o (as T2T and I2T) and DALL-E 3/Stable Diffusion as T2I. I fed the whole Sleepy Hollow story to the LLM, hand-picked five top-imageability sentences extracted from it and asked the LLM to turn each of them into suitable prompts for the T2I part, keeping the story context into consideration (see Appendix A for details). I passed the results to DALL-E 3 and Stable Diffusion (free version), obtaining the images in Fig. 5. It is clear that the LLM has not always succeeded at localizing the prompts (see *the next morning* in prompt 1, the *Dutch country* in prompt 5) and that the T2I systems fail at capturing a coherent atmosphere (though Stable Diffusion is more consistent) and at interpreting the figurative aspects of Ichabod's portrait in Prompt 2 (DALL-E 3 produces a non-human caricature, Stable Diffusion conjures up a bird from the *weathercock* mentioned in the prompt). It should be possible to do better by generating several images, captioning them and selecting those whose captions stay closer to the original text.

The simplest way to tackle the fourth type of I+L output in Fig. 4 — adding text to an image — is to prompt an I2T system not for a caption but for ideas associated with the image given, or thoughts likely to have occurred to the characters in the image. Doing this for the rider image in Fig. 5 leads another LLM (Claude Opus 1.5, unaware of the story the image was generated for) to produce a set of comments such as "*Those crows are unsettling. Is it a bad omen or just my imagination?*", which could be used with the image, but which remain disconnected from the original text (see Appendix B).

This small example suggests that, if we want the text to remain in the driver's seat of the generation process, we need a *splitter*: a module that decides which aspects of the input (including subparts of single sentences) should be cast in graphic form and which ones should continue as text. (2) illustrates some (hand-made) challenging cases.

(2) a. A teddy bear is standing on a skateboard in Times Square, **thinking on how lonely it appeared during COVID peaks.**
b. A teddy bear is standing on a skateboard in Times Square, **because it doesn't have a wallet to pay the cab.**

The splitter should presumably decide that the part in **bold** is best express by language, while the rest can be rendered as an image (see Fig. 7). Broadly speaking, imageability can do part of the work ("a **sinful/nostalgic** teddy bear"), but in other circumstances, what matters is how secondary material is connected to the main clauses ("Teddy is in Times Square **because/when it wants/needs to see people**") or whether the candidate text is an assertion or something else: DALL-E 3 appears to take yes/no questions as instructions to generate ambiguous materials (see Fig. 6) and (like Stable Diffusion) has unpredictable behaviors with wh- questions.[18] Modifiers that embed disjunction or negation ("The teddy bear **or stuffed dog** that **didn't come yesterday** reached Times Square **instead of its brother**, **despite having only a skateboard**") are essentially impossible to render as images alone. In yet other cases, the linguistic relations could be rendered by segmenting the text in multiple images, with modifiers like *before* and *after* used to guide image order.

As mentioned above, one of the problem of an I+L system is image consistency, especially if the image/text ratio is very high. In this case, many contiguous images will contain approximately the same objects, characters and backgrounds. This creates the necessity to have a measure of object consistency. One strategy to this effect is sketched in Fig. 8: a detailed textual description ② is obtained from an original image ① via a pretrained I2T, then used by a pretrained T2I system to generate a derived image ③ (possibly at low-resolution, to be upscaled at the end, see e.g. (Saharia et al. 2022)). The I2T phase could now be fine-tuned using as error function the T-T distance between the description of the original image ② and that of the derived image ④. Alternatively (or concurrently), one could try to minimize a distance metric (I=I) between the original

18 Note that while the pictures on the right in Fig. 6 might strike one as reasonable representations of uncertainty, it is hard to guess from them alone that their prompt was actually a question. Underspecification is everywhere: the image in Fig. 3 doesn't come from "How many delegates were meeting Lincoln?", but it could.

Figure 5
Illustration of 5 Sleepy Hollow sentences, GPT4o-generated prompts from Appendix A. **Top:** DALL-E 3; **Bottom:** Stable Diffusion

Figure 6
Top: "The person with the vase is a woman" (left) "Is the person with the vase a woman?" (right); **Bottom:** "The vase of white flowers was red" (left) "Was the vase of white flowers red?" (right)

Figure 7
A pensive Teddy. Based on a DALL-E 2 image, ©OpenAI

image ① and the derived one ③. This is obviously more difficult, since the same textual description can generate very different images, where "different" does not mean "incompatible" (witness Fig. 1). On the other hand, T=T distance might be oblivious to stylistic differences that the users could find jarring.

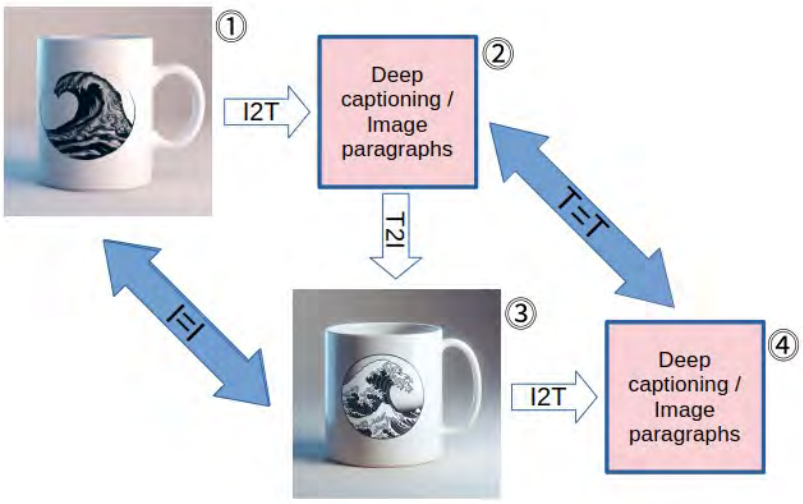

Figure 8
Pipeline schema for exhaustive image descriptions. The oblique arrows (I=I and T=T) represent the distances to minimize via I2T and T2I fine-tuning.

6. The partial information case: rending minimal images

Consider now the situation where an I+L system needs to render a partial text: a dialogue perhaps, or a text presented in steps. Our example will be the passage in (3).

(3) a. In a cozy restaurant there was a table set for two, with wine glasses and a lit candle.
b. A couple arrived and sat at the table.
c. A waiter arrived in a moment and asked which wine they would like: red, white or sparkling?
d. They both took white.

Feeding (3a) alone to DALL-E 3 yields the heart-warming scene in Fig. 9 (the table is not really set for two; let's set this aside). Unfortunately, at (3d) the information comes in that the couple will be drinking *white* wine, not red; now the picture has become inconsistent. Feeding to DALL-E 3 the entire dialogue in (3) does not help (Fig. 10); neither does adding to the prompt requests such as: *Draw only what is explicitely said in this prompt:*, or *Nothing else should be present on or around the table*. Stable Diffusion appears to be better at not showing background fillers, but fails in the number of objects. More generally, current T2I models seem to suffer from a sort of *horror vacui*: more detailed prompts can change the objects, but seem incapable of convincing the model to simply draw less.

The first step in eliminating inconsistent fillers is to identify them. To test the approach in Fig. 8, I experimented with cycling from a description to a set of images, then back to a set of captions, as shown in Fig. 11. I started by generating four versions of an image for the same simple prompt ("*A table set for two, with wine glasses and a lit candle*") using DALL-E 3, then I captioned each image using GPT4. The rationale was that fillers have an important random component and are not likely to be the same across different pictures; the hope was that by asking LLM to collect the *intersection* of the captions we could identify a core containing only the elements from the original prompt, and perhaps identify the most faithful image. Unfortunately, none of the model tested (GPT4, Claude Opus and Gemini 1) came close to the desired results, yielding broad descriptive comments (highlighted in blue) or factually inaccurate descriptions

Figure 9
A table set for one and half

Figure 10
Results of feeding DALL-E 3 the entire romantic dinner dialogue.

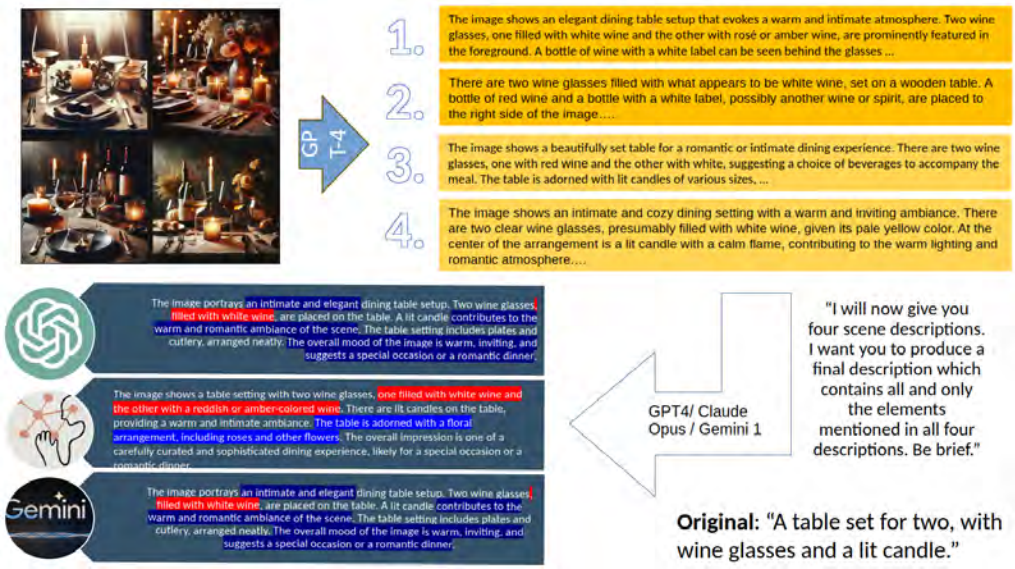

Figure 11
A test to use caption intersection to identify fillers

(in red). Clearly, more work is needed, and experimenting with different I2T models and different captioning engines are all reasonable future directions.

Interactive I+L requires a constant exercise of *reference fixing*: new objects are drawn all the time and may be referred to in the text produced by the system or by the user, even when the image that contained them has been hidden. There seem to be two ways to approach this subtask: one is to search for the objects in the (current and previous) images every time they are mentioned (Fig. 12); the other, to proactively create two versions of each image, one clean, the other with bounding boxes associated with descriptions and unique identifiers (Fig. 13). These objects are stored in a DRT-style storage (Kamp and Reyle 1993), and are matched by subsequent mentions. The advantage of an I+L system is that the answer to a question like *"Is every cat we have encountered so far spotted?"* needs not be a simple *"yes/no"*: the verbal answer can be accompanied by a demonstration, in this case an image that arranges in full sight all the cats seen in the previous images or mentioned in the text (so, not just a collage of "cat"-tagged bounding boxes). This is beyond current capabilities, but not by far.

7. Bridging the gap between still and moving images

The ability of starting from a text or a text+image combination and performing all the operations described so far is not yet a guarantee that we are doing a service to communication, or attention. There is a vast cognitive psychology literature on the relation between language and vision (see (Jackendoff 1987) for a particular linguistic strand, and (Vulchanova et al. 2019)), on how images affect attention and how languages could shape perception. This literature should inform some of the issues touched above: what is the best proportion of language and images for any given genre; what is the best modality to add text to images, etc. But it could also inform new ways of placing images *in time*. I+L systems have in fact the potential of bridging the gap between still images

Figure 12
Finding bounding boxes for the referents for "It" "The bottle", "The center of the table". Image: DALL-E 2

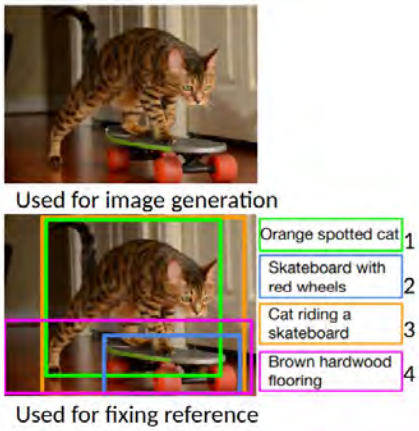

Figure 13
Generating descriptions and numerical identifiers with object bounding boxes

and videos, which is currently only occupied by motion graphics design[19] (i.e. looped animations based on still images, with no or minimal interactivity). Videos are often too fast to convey complex contents in a short time and without annoying stop-back-repeat cycles; still images are normally a-temporal, in the sense that their creator does not know when they will be seen by their intended audience. To be part of an ongoing, on-line discourse, still images need to become sensitive to the flow of the conversation.

Consider for instance *repeated entity individuation*. Language is fond of going back to previously introduced entities using pronouns or definite descriptions, creating chains of reference. A mini-discourse like (4) could be translated into a correct sequence of pictures, but the result would work only if the referents of the boldfaced nouns are clear to the system and evident for the user.

(4) **Two teddy bears** are at a picnic table. **One** decides to chase a butterfly. **The second one** runs after **him**. **A third teddy bear** occupies the picnic table left empty by **them**.

A mixed I+L scenario has the possibility of using graphics to this effect. Let's assume that the initial entities, two teddy bears, are introduced graphically. In referring back to them, the system could accompany a pronoun of a definite description with graphic effects on the image (e.g. making the object mentioned flash or pulse, or blurring/-decoloring the background). The same could happen in reaction to the user's mention: what gets highlighted is what the system thinks the user is referring to, so that errors can be easily spotted and corrected.

Picking up to the same elements in subsequent images could similarly make use of graphic effects. In (4), assuming that the problem of visual consistency has been addressed, the *same* bear returning should be drawn the same way; *different* bears, while recognizably 'beary', must have some element that distinguishes them. This could be accomplished using written tags connected to each object, but style could also be put to use: first-mention could be a full drawing, later mention, a sketched silhouette. Alternatively, contents of which we do not know much, like the romantic

19 See for instance `https://www.zekagraphic.com/how-to-improve-your-website-design-with-motion-graphics/`

table scenario in (3a) above, could be ink sketches with no background; as new details arrive, lines and colors could be added to the image; when all the information has been provided (or, in a dialogue, when all participant have converged on a shared world view) the scenes become oil paintings. In a dialogue, colors or style could be used to code graphic elements that have been provided by different participants, or on which there is disagreement. In a foreign language learning scenario, words and their elements they describe could be color-coded on the fly, to teach the meaning of nouns. As obvious, there will be competition as to which graphic features express which discourse features. Cognitive usability tests will be essential to decide what should take priority.

Other possibilities can be explored starting from the other end — from videos. Consider a video where several people interact in complex ways in a language the viewer don't know well. At present, if something is lost, the viewer can only rewind and play back. Slowing down the video is also a possibility, but it doesn't work well with spoken language. Future I+L systems could associate text lines with the characters in the video (in the form of captions or moving balloons), making the type and amount of text adjust dynamically: slow down the video, and speech is replaced by text. Slow it further, and the text expands to explain who the characters are and what the context is like. Speed the video above normal speed, and you get a textual summary of the broad contents of the conversation. Speed it further, and you get a collection of still images generated from the video, along with a summary of the plot line.

The video-generation side would also change and become more interactive. Asking questions about a point in the video could make the system extract the characters and let the user interact with them before resuming the main storyline. Splitting a video is two windows could be used to render alternative possible developments the user can chose from. With improvements in video generation and object consistency, the limits to what I+L generation could do are our imagination, and our ability to parse multimodal signals efficiently and effortlessly. And if future AI systems start to communicate among themselves using multimodal channels, there is no reason why their exchange should stop at text and images: AIs don't have a pre-set number of senses, and could learn to interact, for instance, by exchanging their own internal states (whatever their use or meaning). Their limit is not modality, only interpretability.

8. Conclusions

This opinion piece is a call to arms for a new family of AI research tasks, the generation of mixed languages and images, which is starting to become possible due to advances in neighboring tasks: text-to-text, image-to-image, text-to-image, image-to-text (in captioning), text+image-to-image (in "guided" language generation), text+image-to-text (in VQA). The missing element, text(+image)-to-text+image, is not absent by chance. While potentially useful (we do use images in public presentations, we do have illustrated books and comics), it is a combination that our species cannot produce at "on-line" speed: a speed sufficient for dialogue. The consequence is that few if any of the many datasets accumulated by AI researchers can be directly used to train it or test it, though some (instructional videos, comics, film+storyline) could be the base to create useful data. The creation of illustrated narrative texts is a first step, and I have suggested that the output of T2T generation system (or more image-ready language variants) could be preprocessed in a way that separates the parts that can be converted into images or image sequences from those that cannot. This could be carried out by a (looped) combination of current I2T and T2I generators, to be verified by humans until a critical mass of training and testing materials has accumulated. To conclude, mixing

language and images in a complementary way raises a number of interesting theoretical and applied issues for computational linguistics and cognitive science. Embracing this innovation requires addressing them with a open mind, and seriously considering what it means to create datasets that go beyond human abilities.

Acknowledgments

Many thanks to Dominique Brunato, Alessandro Lenci and the audiences of talks at the NL4AI@AIxIA 2023 workshop and the university of Pisa for insightful discussions. Special thanks to the reviewers of this Special Issue, which pointed out limits and possible expansions. All errors and inaccuracies are of course my own.

References

An, Jie, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Lijuan Wang, and Jiebo Luo. 2024. Openleaf: A novel benchmark for open-domain interleaved image-text generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 11137–11145, New York, NY, USA, October. Association for Computing Machinery.

Antol, Stanislaw, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, Lawrence C. Zitnick, and Devi Parikh. 2015. VQA: visual question answering. *CoRR*, abs/1505.00468.

Arijon, Daniel. 1976. *Grammar of the Film Language*. Silman-James Press.

Barthes, Roland. 1964. Rhetoric of the image. In Roland Barthes, editor, *Image–Music–Text*. Fontana, London, pages 32–51. Reprinted in 1997.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In Hugo Larochelle, M. Ranzato, Raya T. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Cohn, Neil. 2013. *The Visual Language of Comics. Introduction to the Structure and Cognition of Sequential Images*. Bloomsbury.

Della Rosa, Pasquale A., Eleonora Catricalà, Gabriela Vigliocco, and Stefano F. Cappa. 2010. Beyond the abstract–concrete dichotomy: Mode of acquisition, concreteness, imageability, familiarity, age of acquisition, context availability, and abstractness norms for a set of 417 Italian words. *Behavior Research Methods*, 42(4):1042–1048.

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, Miami, June.

Dong, Runpei, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, Xiangwen Kong, Xiangyu Zhang, Kaisheng Ma, and Li Yi. 2024. DreamLLM: Synergistic multimodal comprehension and creation. In *The Twelfth International Conference on Learning Representations*, Wien, May.

Evans, Talfan, Nikhil Parthasarathy, Hamza Merzic, and Olivier J Henaff. 2024. Data curation via joint example selection further accelerates multimodal learning. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, Vancouver, December.

Fernandes, Daniel L., Marcos H. F. Ribeiro., Fabio R. Cerqueira., and Michel M. Silva. 2022. Describing image focused in cognitive and visual details for visually impaired people: An approach to generating inclusive paragraphs. In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP,*, pages 526–534, Virtual Event, February. INSTICC, SciTePress.

Gal, Rinon, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. 2023. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, Kigali, May.

Ghandi, Taraneh, Hamidreza Pourreza, and Hamidreza Mahyar. 2023. Deep learning approaches on image captioning: A review. *ACM Computing Surveys*, 56(3), October.

He, Yingqing, Zhaoyang Liu, Jingye Chen, Zeyue Tian, Hongyu Liu, Xiaowei Chi, Runtao Liu, Ruibin Yuan, Yazhou Xing, Wenhai Wang, Jifeng Dai, Yong Zhang, Wei Xue, Qifeng Liu, Yike Guo, and Qifeng Chen. 2024. LLMs meet multimodal generation and editing: A survey. *CoRR*, abs/2405.19334.

Jackendoff, Ray. 1987. On beyond zebra. *Cognition*, 26(2):89–114.

Johnson, Justin, Andrej Karpathy, and Li Fei-Fei. 2016. DenseCap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, June.

Kamp, Hans. and Uwe Reyle. 1993. *From Discourse to Logic*. D. Reidel, Dordrecht.

Kar, Sudipta, Suraj Maharjan, A. Pastor López-Monroy, and Thamar Solorio. 2018. MPST: A corpus of movie plot synopses with tags. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Kawar, Bahjat, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2023. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6007–6017, Vancouver, June.

Krause, Jonathan, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. 2017. A hierarchical approach for generating descriptive image paragraphs. In *Computer Vision and Patterm Recognition (CVPR)*, pages 317–325, Honolulu, July.

Kucherenko, Taras, Pieter Wolfert, Youngwoo Yoon, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2024. Evaluating gesture generation in a large-scale open challenge: The GENEA challenge 2022. *ACM Transactions on Graphics*, 43(3), June.

Leitch, Thomas. 2007. *Film Adaptation and Its Discontents: From Gone with the Wind to The Passion of the Christ.* Project MUSE. Johns Hopkins University Press.

Li, Shenggui, Hongxin Liu, Zhengda Bian, Jiarui Fang, Haichen Huang, Yuliang Liu, Boxiang Wang, and Yang You. 2023. Colossal-AI: A unified deep learning system for large-scale parallel training. In *Proceedings of the 52nd International Conference on Parallel Processing*, ICPP '23, page 766–775, Salt Lake City, Utah, USA, August. Association for Computing Machinery.

Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*. Springer International Publishing, Cham, pages 740–755.

Lindfors, Hanna, Kristina Hansson, Eric Pakulak, Neil Cohn, and Annika Andersson. 2024. Semantic processing of verbal narratives compared to semantic processing of visual narratives: an ERP study of school-aged children. *Frontiers in Psychology*, 14.

Liu, Heng, Yao Xu, and Feng Chen. 2023. Sketch2photo: Synthesizing photo-realistic images from sketches via global contexts. *Engineering Applications of Artificial Intelligence*, 117:105608.

Liu, Yu, Gelareh Mohammadi, Yang Song, and Wafa Johal. 2021. Speech-based gesture generation for robots and embodied agents: A scoping review. In *Proceedings of the 9th International Conference on Human-Agent Interaction*, HAI '21, page 31–38, Virtual Event Japan, November. Association for Computing Machinery.

Ljubešić, Nikola, Darja Fišer, and Anita Peti-Stantić. 2018. Predicting concreteness and imageability of words within and across languages via word embeddings. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 217–222, Melbourne, Australia, July. Association for Computational Linguistics.

Lu, Xiaopeng, Lynnette Ng, Jared Fernandez, and Hao Zhu. 2021. CIGLI: conditional image generation from language & image. *CoRR*, abs/2108.08955.

Ma, Yan, Yu Qiao, and Pengfei Liu. 2024. MoPS: Modular story premise synthesis for open-ended automatic story generation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2135–2169, Bangkok, Thailand, August. Association for Computational Linguistics.

Martinec, Radan and Andrew Salway. 2005. A system for image–text relations in new (and old) media. *Visual Communication*, 4(3):337–371.

McNeill, David. 1992. *Hand and Mind: What Gestures Reveal About Thought*. The University of Chicago Press.

Mildenhall, Ben, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2021. Nerf: representing scenes as neural radiance fields for view synthesis.

Communications of ACM, 65(1):99–106, December.

Nichol, Alexander Quinn, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 16784–16804, Baltimore, 17–23 July. PMLR.

Palma, Tobias, Guy Schofield, and Grace Moore. 2023. Narrative perspectives and embodiment in cinematic virtual reality. In Lucio T. De Paolis, Pasquale Arpaia, and Marco Sacco, editors, *XR Salento 2023 – International conference on aXtended Reality*. Springer, Lecce, September, pages 232–252.

Pezzelle, Sandro, Claudio Greco, Greta Gandolfi, Eleonora Gualdoni, and Raffaella Bernardi. 2020. Be Different to Be Better! A Benchmark to Leverage the Complementarity of Language and Vision. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2751–2767, Online, November. Association for Computational Linguistics.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 8748–8763, Virtual Event, 18–24 Jul. PMLR.

Ramesh, Aditya, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with CLIP latents. *arxiv*, abs/2204.06125.

Ruiz, Nataniel, 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*, Vancouver, June.

Saharia, Chitwan, William Chan, Saurabh Saxena, Lala Lit, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Raphael Gontijo-Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. 2022. Photorealistic text-to-image diffusion models with deep language understanding. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, November. Curran Associates Inc.

Shen, Sheng, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. 2022. How much can CLIP benefit vision-and-language tasks? In *2022 International Conference on Learning Representations*, Virtual Event, April.

Spanjers, Rik. 2021. 3 Text-image relations. In Sebastian Domsch, Dan Hassler-Forest, and Dirk Vanderbeke, editors, *Handbook of Comics and Graphic Narratives*. De Gruyter, Berlin, Boston, pages 81–98.

Tewel, Yoad, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. 2024. Training-free consistent text-to-image generation. *ACM Transactions on Graphics*, 43(4), July.

Vulchanova, Mila, Valentin Vulchanov, Isabella Fritz, and Evelyn A. Milburn. 2019. Language and perception: Introduction to the special issue "Speakers and listeners in the visual world". *Journal of Cultural Cognitive Science*, 3:103–112.

Wolfert, Pieter, Nicole Robinson, and Tony Belpaeme. 2022. A review of evaluation practices of gesture generation in embodied conversational agents. *IEEE Transactions on Human-Machine Systems*, 52(3):379–389.

Yu, Lili, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, Candace Ross, Adam Polyak, Russell Howes, Vasu Sharma, Puxin Xu, Hovhannes Tamoyan, Oron Ashual, Uriel Singer, Shang-Wen Li, Susan Zhang, Richard James, Gargi Ghosh, Yaniv Taigman, Maryam Fazel-Zarandi, Asli Celikyilmaz, Luke Zettlemoyer, and Armen Aghajanyan. 2023. Scaling autoregressive multi-modal models: Pretraining and instruction tuning. *arXiv*, abs/2309.02591.

Zamparelli, Roberto. 2023. One picture and a thousand words. generative language+images models and how to train them. In Elisa Bassignana, Dominique Brunato, Marco Polignano, and Alan Ramponi, editors, *Proceedings of the Seventh Workshop on Natural Language for Artificial Intelligence (NL4AI 2023) co-located with 22th International Conference of the Italian Association for Artificial Intelligence (AIxIA 2023), Rome, Italy, November 6th-7th, 2023*, volume 3551 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Zhou, Yufan, Ruiyi Zhang, Jiuxiang Gu, Chris Tensmeyer, Tong Yu, Changyou Chen, Jinhui Xu, and Tong Sun. 2022. Tigan: Text-based interactive image generation and manipulation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3):3580–3588, Jun.

Appendix A: Prompt to GPT4o (see Fig: 4)

"Now I will upload the text of a short story. I will then give you five sentences, extracted from the short story; you task will be to turn each of them into a prompt to generate an image that illustrates the story. The images should all have the same style, and should be consistent with the time and location of the story, so each prompt must contain enough information to achieve this goal. Ready?"

INPUT HIGH-IMAGEABILITY SENTENCES:

1. The next morning the old horse was found without his saddle, and with the bridle under his feet, soberly cropping the grass at his master's gate.
2. His head was small, and flat at top, with huge ears, large green glassy eyes, and a long snipe nose, so that it looked like a weather-cock perched upon his spindle neck to tell which way the wind blew.
3. Away, then, they dashed through thick and thin; stones flying and sparks flashing at every bound.
4. On that side of the road where the brook entered the wood, a group of oaks and chestnuts, matted thick with wild grape-vines, threw a cavernous gloom over it.
5. There was the doughty doughnut, the tender oly koek, and the crisp and crumbling cruller; sweet cakes and short cakes, ginger cakes and honey cakes, and the whole family of cakes.

OUTPUT:

1. Illustration Prompt:
   "Depict the scene of the next morning where an old horse is found without his saddle and with the bridle under his feet. The horse is calmly grazing on the grass at his master's gate in a rural 18th-century setting. The scene should capture the quiet and serene morning atmosphere, with colonial-style buildings and natural surroundings."
2. Illustration Prompt:
   "Create a detailed character portrait of Ichabod Crane. He should be depicted with a small, flat-topped head, huge ears, large green glassy eyes, and a long snipe nose, resembling a weather-cock perched on a spindle neck. The setting should be consistent with the early American, colonial era, capturing the eerie yet quaint atmosphere of Sleepy Hollow."
3. Illustration Prompt:
   "Illustrate an intense action scene where Ichabod Crane dashes through the dark, spooky forest on horseback, with his steed galloping at full speed. Show stones flying and sparks flashing at every bound,

emphasizing the urgency and fear in the night-time setting typical of Sleepy Hollow."

4. Illustration Prompt:
"Depict the gloomy and cavernous entrance to the woods where brook meets the road. A dense group of oaks and chestnuts, entwined with wild grapevines, should create a dark, shadowy atmosphere. The scene should evoke a sense of foreboding and mystery, consistent with the haunted landscape of Sleepy Hollow."

5. Illustration Prompt:
"Create a vibrant and detailed illustration of a sumptuous Dutch country tea-table laden with various cakes. Include the doughty doughnut, tender oly koek, crisp and crumbling cruller, sweet cakes, short cakes, ginger cakes, and honey cakes. Capture the rustic charm and abundance typical of a Dutch farmhouse in the 18th century, with warm lighting and homely atmosphere."

Appendix B: Generating non-graphic comments from the rider's image

PROMPT: What thoughts are likely to go through this rider's mind?

ANTHROPIC'S CLAUDE OPUS 1.5 (UNAWARE OF THE STORY FOR WHICH THE IMAGE WAS GENERATED):

Based on the intense and urgent scene depicted in the image, the rider's thoughts are likely to be focused and driven. Some possible thoughts going through the rider's mind could include:

- "I've got to hurry - there's no time to lose!"
- "Stay focused. Don't let this eerie forest distract me from my mission."
- "Come on, boy! Faster! We're almost there!"
- "What dangers lie ahead? I must be prepared for anything."
- "I hope I'm not too late. The fate of others may depend on me."
- "Keep steady. Don't let the rough terrain slow us down."
- "Those crows are unsettling. Is it a bad omen or just my imagination?"
- "I can't fail now. Too much is at stake."
- "The town lights are getting closer. Just a little further!"
- "Stay calm. Control your fear. Focus on the task at hand."

The rider appears to be in a state of high alertness and determination, likely driven by a sense of urgency or importance in whatever mission or task he's undertaking. The ominous setting suggests he might also be experiencing some level of apprehension or concern about potential dangers, while still remaining resolute in his purpose.

OPENAI'S GPT4O (AWARE OF THE STORY FOR WHICH THE IMAGE WAS GENERATED)

Given the intense and frightening scene depicted, here are some thoughts that are likely to go through the rider's mind:

Fear and Panic: The rider is probably overwhelmed with fear due to the alarming presence behind him and the ominous, dark forest setting.

- "What is that behind me? Is it the Headless Horseman?"
- "I need to escape this place immediately!"

Urgency and Desperation: The rider is clearly in a rush, urging his horse to move faster to outrun whatever is pursuing him.

- "Come on! Faster, faster!"
- "If I can just make it to the bridge, I'll be safe!"

Confusion and Disbelief: The supernatural elements and the eerie environment may make the rider question what he is experiencing.

- "Is this really happening? Am I imagining things?"
- "How can this be real?"

Hope and Determination: Amidst the fear, there's likely a sense of hope and determination to survive the ordeal.

- "I can make it through this. Just a little bit further."
- "I must stay focused and not let fear take over."

Reflection and Regret: The rider might also reflect on the events leading up to this moment, possibly feeling regret.

- "Why did I leave so late?"
- "I should have listened to the warnings about this place."

These thoughts create a complex mixture of emotions that would be coursing through the mind of someone in such a perilous situation.