Symposium: Repurposing Generative AI for Social Research – *peer-reviewed*
*Edited by Federico Pilati, Anders Kristian Munk, and Tommaso Venturini*
https://doi.org/10.6092/issn.1971-8853/19512

Sociologica. V.18 N.2 (2024)
ISSN 1971-8853
https://sociologica.unibo.it/

# Synthetic Probes: A Qualitative Experiment in Latent Space Exploration

Gabriele de Seta*

Department of Linguistic, Literary and Aesthetic Studies, University of Bergen (Norway)

Submitted: May 9, 2024 – Revised version: September 9, 2024
Accepted: September 23, 2024 – Published: October 30, 2024

## Abstract

This essay outlines a methodological approach for the qualitative study of generative artificial intelligence models. After introducing the epistemological challenges faced by users of generative models, I argue that these black-boxed systems can be explored through indirect ways of knowing what happens inside them. Inspired by both ethnographic and digital methods, I propose the use of what I call *synthetic probes*: qualitative research devices designed to correlate the inputs and outputs of generative models and thus gather insights into their training data, informational representation, and capability for synthesis. I start by describing the sociotechnical context of a specific text-to-video generative model (ModelScopeT2V), and then explain how my encounter with it resulted in an extensive period of experimentation dedicated to the production of *Latent China*, a documentary entirely composed of synthetic video clips. Reflecting on how this experience bridges qualitative research and creative practice, I extrapolate more general observations about how a long history of research probes across disciplines can inspire the creation of methodological devices designed to allow the indirect exploration of a machine learning model's latent space.

**Keywords**: Generative artificial intelligence; latent space; machine learning models; probes; qualitative methods.

## Acknowledgements

The supporting agency for this contribution is Trond Mohn Foundation, grant no. TMS2024STG03

* ✉ gabriele.seta@uib.no

## 1 Probes in Latent Space

Besides classifying data and extrapolating predictions, machine learning models are increasingly used to generate information. The domain of machine learning commonly called "generative artificial intelligence" encompasses models designed to synthesize new text, images, sounds, or other kinds of content according to the datasets they have been trained on. Through a computationally intense process of training, the model "learns" to represent a dataset in a more abstract form as a collection of high-dimensional vectors called "latent space". Once trained, generative models synthesize information according to inputs including random seeds, user prompts, guidance images, as well as parameters such as temperature or inference steps, which all influence the resulting output. Any single output generated by one of these models consists of a minuscule slice of this latent space, reduced to a manageable number of dimensions for human (or machine) interpretation. While generative models are ultimately deterministic — i.e., the same combination of random seed, prompt, and input parameters result in the same output — the scale and complexity of their computational architectures makes them opaque to human interpretation. Even if one could observe the exact configuration of weights determining the response generated by a large language model (LLM), the array of pixels synthesized by a text-to-image model, or the waveform produced by a text-to-speech model, it would be impossible to extrapolate what exactly the model had generalized from the training data and how this contributed to its output. For users of generative models, who might not have direct access to neither the model itself nor the training data, this entails an epistemological challenge, as all that is available for interpretation is the input and the output, with everything in between hidden away inside nested black boxes.

Following Bill Maurer's methodological metaphor, this essay proposes that these nested black boxes can be, if not opened and examined, at least shaken for clues about their functioning (Ziewitz, 2016). My argument is that, while the high-dimensional nature of latent spaces makes them fundamentally impenetrable to human cognition, the correlation between inputs and outputs can be operationalized to obtain some insights into the data a model has been trained on, what the model has learned from it, and how the model draws upon it to synthesize new information. As a qualitative researcher, I approach these questions from the perspective of everyday use at the human scale. By combining the contextual and dialogic depth afforded by ethnographic research with the digital methods intuition of studying a medium through the medium itself (Rogers, 2013), I propose to experiment with the creation of methodological devices that allow the indirect exploration of a machine learning model's latent space. In the following section, I begin with an ethnographic *entrée* into the field by introducing the sociotechnical context of a specific machine learning model, Alibaba's ModelScopeT2V, highlighting the situated and contingent development of artificial intelligence as a bricolage of platforms, tools, and interfaces. Section 3 begins from my encounter with ModelScope T2V and describes the creative process behind *Latent China*, a synthetic documentary entirely composed of footage generated by the model when asked to represent its country of origin. In section 4, I draw on this experiment to generalize a methodological approach that can be used to explore the latent spaces of generative models: the development of research devices which, inspired by the use of probes in ethnographic and design research, I call "synthetic probes". In the conclusion, I offer some more general observations for opening up new research trajectories for these probes amidst the proliferation of machine learning models, automated agents and algorithmic systems.

## 2 Of Scopes and Models

On November 3, 2022, Alibaba Cloud (the AI and cloud computing subsidiary of Chinese conglomerate Alibaba Group) unveiled a new open-source MaaS (Model-as-a-service) platform called ModelScope, comparable to machine learning platforms such as Hugging Face or Azure (Gan et al., 2023). According to Jeff Zhang, President of Alibaba Cloud, this platform was part of an effort to "lower the barrier for companies to adopt new technology and capture more opportunities in the cloud era" (Alibaba Cloud Community, 2022). At launch, ModelScope featured more than 300 AI models developed by Alibaba's own research unit, DAMO Academy, which offered tasks such as computer vision, natural language processing and image captioning. ModelScope's press release emphasized its commitment to open-source computing and community support:

> Developers and researchers can simply test the models online for free and get the results of their tests within minutes. They can also develop customized AI applications by fine-tuning existing models, and run the models online backed by Alibaba Cloud, or deploy them on other cloud platforms or in a local setting.

Nine months after its launch, Alibaba's vice president Ye Jieping claimed that ModelScope hosted over two million users and more than 1,000 large models, including open-source ones from both Chinese firms and foreign ones (TechNode Feed, 2023). By November 2023, the total number of models reached 2,300, including Alibaba's own large language model Tongyi Qianwen, and ModelScope had arguably become the largest AI model community in China (Yu, 2023).

The ModelScope platform is web-based, and its homepage welcomes users with a minimalist interface: a large button invites to "enter the community area" through an onboarding login with the most recent commercial promotion; on the right side of the page, a scroll-down list presents the most popular models and datasets, which at the time of writing include Alibaba's own Qwen 1.5. and Meta's llama-3 and MusicGen. The website is divided into a few main sections: Models, Datasets, Creator Space, Documentation Center, and Communities. The Models section allows users to explore the 4,548 machine learning models available at the time of writing by popularity, language, or type (computer vision, NLP, voice, multimodal, scientific calculation) and to access documentation, demos and codebases. For example, one of the most popular text-to-video synthesis models is ModelScopeT2V, uploaded by Alibaba's own Tongyi Lab on 21 March 2023 and last updated on 30 November 2023, which has been downloaded more than a hundred thousand times and has received more than 500 likes. The web page dedicated to this specific model describes it as a "multi-stage text-to-video generation diffusion model" which generates a video output according to a descriptive text input through the iterative denoising of pure Gaussian noise. As the description briefly explains, ModelScopeT2V is in itself a combination of three sub-networks ("text feature extraction, text feature-to-video latent space diffusion model, and video latent space to video visual space") trained on multiple datasets and totaling 1.7 billion parameters (Institute for Intelligent Computing, 2023).

Some output examples provided in the model description page include short video clips prompted by sentences like "robot dancing in times square," "a cat eating food out of a bowl, in the style of van Gogh" and "balloon full of water exploding in extreme slow motion." In the "How to use" section, users are invited to test the model on either the ModelScope Studio or the Hugging Face platforms, or to refer to a notebook tutorial and set up their own implementation. A short section on limitations and biases highlights the model's restriction to English language prompts and warns about the model's training on public datasets which might skew its outputs (it cannot generate neither film and TV-quality video nor text). A similar section on misuse warns against commercial, demeaning, harmful, pornographic, and deceptive uses of the model: while the output examples seem to highlight realism and accuracy (the model is even tagged with "realism"), a disclaimer reads: "The model was not trained to realistically represent people or events, so using it to generate such content is beyond the model's capabilities." In the technical report written by Alibaba researchers that is referenced on the same page, ModelScopeT2V is described in more detail as an evolution of the Stable Diffusion text-to-image model which brings two technical innovations to the field of video generation: a spatio-temporal block to improve consistency and a multi-frame training strategy leveraging multiple datasets (Wang et al., 2023, p. 1). The model combines elements of other generative models (for example, a CLIP text encoder, the VQGAN encoder/decoder, and a denoising UNet) to achieve the diffusion-based synthesis of videos from a latent space to a visual space (p. 3). In line with this narrative of incremental innovation, the authors present ModelscopeT2V as a publicly available platform for further innovations in video generation.

In comparison with other thousands of models uploaded on the platform, ModelScopeT2V managed to reach a rather wide audience: as soon as a day after its release, a Gizmodo report hailed it as "the first AI video generator to catch the internet's attention," claiming that "text to video generative AI is finally here and it's weird as hell" (Barr, 2023). Being released slightly before competitors like Runway, Google or Meta were able to showcase their text-to-video capabilities, the model developed by DAMO Academy gave many everyday users their first chance to play around with generative video: "The internet is freaking out over AI-generated videos that are so bad you can't look away," a Business Insider article on the model reported in late March (Mok, 2023). The model's popularity owes much to a compilation of video outputs created by Reddit user chaindrop with the prompt "Will Smith eating spaghetti," which depicted the star in weird and uncanny interactions with pasta (chaindrop, 2023). As reported by popular tech outlets, the Will Smith eating spaghetti meme propelled the ModelScope text-to-video model into worldwide popularity (Cole, 2023), leading other users to generate their own short clip compilations of other absurd subjects, such as Joe Rogan fighting a bear, Dwayne "the Rock" Johnson eating rocks, or Elon Musk fighting robots (Hoover, 2023). Thanks to these viral outputs, the ModelScopeT2V became one of the few generative AI tools such as DALL-E, ChatGPT, MidJourney, Stable Diffusion or Suno which the general public would recognize — if not from its name, at least from outputs like the Will Smith eating spaghetti footage, which the star himself made fun of in February 2024, by filming himself eating pasta in weird and unsettling ways (Figure 1).

This brief walkthrough, which started from a massive machine learning platform established by a Chinese tech company and zoomed into one specific model — focusing on its background, functioning and popular culture afterlife — is meant to emphasize some important contextual aspects of these tools: behind generalizations about AI are countless models and datasets with domain-specific capabilities and limitations; models are often bricolages of other models and systems aiming at incremental advancements in narrow tasks; and the use cases envisioned by model creators are not always in line with how broader communities of users adopt them. These observations also have substantial implications for research. As qualitative researchers across disciplines debate how to best integrate artificial intelligence in their methodological pipelines as both tool and collaborator (Jiang et al., 2021) while also worrying about ethical challenges (Davison et al., 2024), it is increasingly important to develop methods to situate, disaggregate, explore and analyze the functioning of specific tools, models, and systems

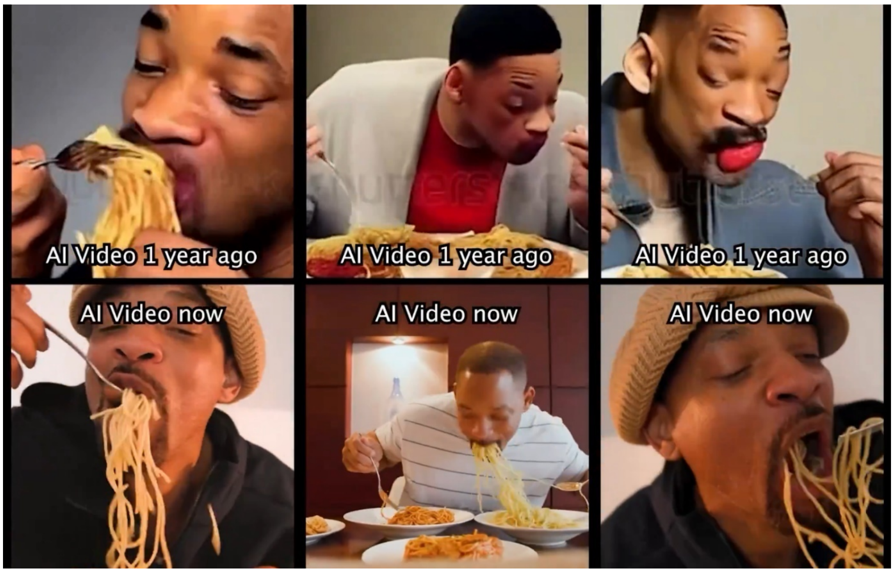

Figure 1. The reaction video posted by Will Smith, in which he re-enacts chaindrop's ModelScopeT2V compilation to parody AI improvement narratives (Will Smith, 2024).

(Elish & boyd, 2018). Computer scientists, data scientists, and computational social scientists already do this extensively through quantitative studies, big data analytics and machine learning itself (Wang et al., 2024) — there is no reason not to expand these efforts through qualitative research sensitized to the methods of these models. After all, on both the ModelScope and the Hugging Face platforms, a disclaimer reminds users that ModelScopeT2V is "meant for research purposes" and not for commercial ones. This article takes the invitation seriously and devises a way to conduct qualitative research about a machine learning model through the model itself.

## 3 Latent China: An Experiment

In late March 2023, like many other people around the world, I started playing around with ModelScopeT2V. My go-to implementation was the one uploaded by the Alibaba TongYi Vision Intelligence Lab to French-American platform Hugging Face, which offered a quite limited interface and at times extenuating waiting times, but had the advantage of being accessible via a web browser from any device. On Hugging Face, the model can be used to generate a single video at a time by inputting a prompt in a text box. Only a few parameters can be tweaked in the "Advanced options": a random seed, the number of frames (limited to a maximum of 32), and the number of inference steps (10 to 50). The model outputs short video clips of up to four seconds with a square resolution of 256 by 256 pixels, a formal constraint that seems to be directly related to decisions made during training: ModelScopeT2V has been trained on a selection of video-text pairs from the WebVid dataset, trimmed down to their middle square portion and sampled for a random subset of 16 frames (Wang et al., 2023, p. 6), and its ability to maintain temporal consistency might be limited at longer lengths. The quite restrictive output format explains the emergence of the compilation as a creative strategy developed by ModelScopeT2V users to offset the limited length of clips by combining several together into longer videos such as the one of Will Smith eating spaghetti.

After testing prompts of different kinds and creating my own share of humorous and absurd content to share on social media, I set forth to explore ModelScopeT2V in a more structured and systematic way. In contrast to most other popular generative models and tools such as Stable Diffusion, ChatGPT or Suno, ModelScopeT2V was developed and released by a Chinese tech company; given my long-standing interest in China's digital development, I decided to pose a rather straightforward question: how does this flagship model uploaded on the largest Chinese machine learning platform represent its country of origin? In order to answer this question, I started prompting the model with very simple, minimal prompts such as "China" or "Chinese" — interestingly, as the model disclaimer explains, ModelScopeT2V is also trained on the LAION2B-en subset and can thus only interpret English-language prompts. My first results were underwhelming: more than half of the outputs were undulating patterns of blobs or stripes, clearly overfitting the starting seeds of Gaussian noise into meaningless abstractions; the other half were more representational, and yet quite random, ranging from blurry metropolitan sights to formless geographical maps. The prompt was too vague to draw any conclusions, and at most demonstrated how the model's VQGAN encoder/decoder module pulled video frames out of the latent space resulting from its training process according to the combination of textual tokens and random seeds.

In search of a more productive process, I started adding nouns to the terms "China" and "Chinese", testing prompts like "Chinese person", "Chinese landscape", or "Beijing, China". Results were more consistent, but the short length of the video output made it difficult to gain any substantive insights into the model's representational range. To offset this limitation, I started generating multiple outputs from the same prompt — first only five or six, then ten, twenty or even thirty, repeating the process until I felt like the outputs exhausted the range of combinatorial possibilities resulting from a specific sequence of tokens. Some prompts, like "Chinese architecture", turned out to be quite productive and interesting, consistently resulting in clearly identifiable outputs with a wide variety of visual content. Others, such as "Chinese internet cafe" or "future China", led me into noisy and abstracted dead ends, being perhaps too specific or lacking representation in the training data to generate any meaningful output. Small variations in wording appeared to correlate to substantial content variations: for example, the prompt "Chinese man" consistently generated long shots of featureless male figures walking on gray pavements, while "man, China" resulted in more dynamic, cinematic and colorful scenes featuring close-ups of clearly Asian men. Accumulating outputs also started revealing patterns and trends that would have been difficult to extrapolate from viewing one clip at a time, including color palettes, subjects, camera angles and camera movements (Figure 2).

Over six months, I generated more than 1,000 China-related 4-second clips, extending my approach to over 80 different prompts. In contrast to the maximalist, detail-oriented guidelines recommended by "prompt engineering" tutorials, my approach to ModelScopeT2V sought to map out a minimal ontology of categories related to China that were narrow enough to produce recognizable outputs, but also broad enough to reveal something about the model and its training rather than my own request. For example, I wanted to see how the "Great Wall of China" was depicted when no further guidance was provided by the prompt. As it turns out, the answer is: quite consistently. All of the clips generated by the model depict sections of the Great Wall on brownish-green mountain slopes or hillsides from a distant point of view, perhaps that of a tourist with a telephoto lens or a filmmaker on a helicopter. Similarly to most other ModelScopeT2V outputs, these depictions of the Great Wall are not entirely realistic, as sections of the architectural marvel move around, split and merge with one another in the span of a few seconds. By testing semantically adjacent prompts and comparing samples of their out-

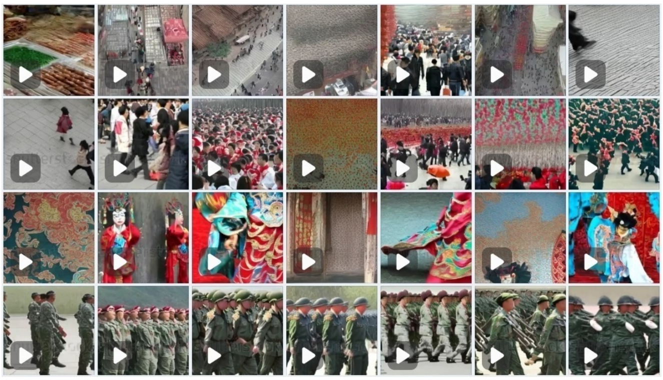

Figure 2. Selection of clips generated with ModelScopeT2V highlighting the shared aesthetic features resulting from prompts such as "Chinese people" (crowded streets, overhead views), "Chinese festival" (dense crowds, red lanterns or clothing), "Chinese opera" (close-ups of floral patterns and traditional costumes), or "Chinese army" (side tracking shots of marching soldiers in green uniforms).

puts that achieved a satisfactory degree of aesthetic saturation, I started seeing some underlying patterns. The prompt "Chinese landscape" largely results in aerial views of mountainous landscapes covered in trees; "Chinese village" maintains this as a backdrop, but adds rather blurred clusters of brown wooden houses and sometimes mountaintop pavilions; "Chinese city" replaces mountains and forests with unstable grids of streets and buildings, but the point of view remains consistently airborne.

All the 1,000 ModelScopeT2V clips I generated for this experiment share one glaringly obvious feature: a Shutterstock watermark — or rather, more precisely, a quite accurate approximation of a watermark generated by the same diffusion process that synthesizes every frame of the output, centered in the bottom half of the video square. This phenomenon has been consistently observed since the release of the model, and while the ModelScopeT2V page does not mention it directly, many have identified its origin in the training process. The "multi-frame training approach" pioneered by the creators of ModelScopeT2V means that, in practice, the model was trained on both a video-text paired dataset (WebVid) and a much larger image-text paired dataset (LAION5B) meant to complement the much smaller scale of the video one, as "training solely on video-text paired datasets can hinder semantic diversity and lead to catastrophic forgetting of image-domain expertise during training" (Wang et al., 2023, p. 5). Since WebVid is a dataset compiling ten million video previews from the stock footage platform Shutterstock, alongside their captions such as "writing robot arm on display at a technology fair in Shanghai," "view from luhuitou park on hainan island, waves approaching the shore," or "young beautiful asian woman home alone watching television smiling laughing China," it is not surprising that ModelScopeT2V had learned that the Shutterstock watermark is the most defining feature of its training data, and reproduced it in nearly any output regardless of its content.

Watching through hundreds and hundreds of 4-second clips gradually allowed me to observe a number of emerging phenomena, some of which are in line with what researchers have observed in the outputs of other machine learning models. Certain prompts, like "Chinese couple" and "Chinese family" unfailingly produced clips of heteronormative couples and nuclear families, a form of naturalization that perpetuates the representational biases encoded in datasets (Denton et al., 2021). Other prompts, like "Mao Zedong" or "Tiananmen Square

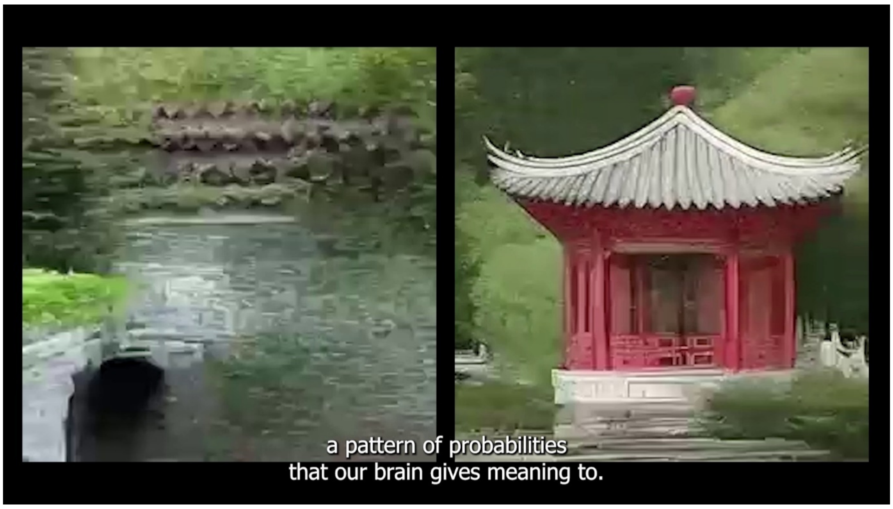

Figure 3. Screenshot from a section of *Latent China* composed of clips resulting from the prompt "Chinese garden".

protest"[1], generated clips reproduced the aesthetics of black and white film or grainy analog television, foregrounding the historical connection between historical personalities or events and the material temporality of media technologies (Offert, 2023). The outputs of some prompts were interesting only when compared: the names specific architectural landmarks such as "Forbidden City" or "Temple of Heaven" generated rather stable and accurate depictions from fixed points of view, while a more general one like "Chinese architecture" often defaulted to a front view of a wooden temple or palace facade. Names of different Chinese cities correlated with specific visual features — Beijing (building walls), Shanghai (nighttime skyscrapers), Hangzhou (water surface), Chongqing (tall buildings between steep hills), Kashgar (beige, sandy streets); names of China's ethnic minorities (Tibetans, Uyghurs, Mongolian) correlated to stereotyped minority clothing and colors. The more I generated and watched ModelScopeT2V clips, the more I felt like I was looking through some kind of optical instrument — a sort of a scope, appropriately — into a latent space abstracting millions of seconds of Shutterstock material shot by photographers and video makers from all over the world. This realization drove my decision to compile these outputs into *Latent China*, a synthetic documentary made not with stock footage, but with its machine-learned approximations generated by one of the first text-to-video models to gain worldwide popularity.

## 4 Synthetic Probes

While composed almost entirely of synthetic content, *Latent China* is by no means a movie made by artificial intelligence — to the contrary, its final version is the result of many authorial decisions and creative processes, including extensive periods of categorization, selection, sequencing and editing work. These also include the choice of the video compilation format, inspired by the vernacular adoption of ModelScopeT2V to generate short humorous videos, as well as its framing in the documentary genre, driven by my realization of the predominance of

[1] It is important to note that the Hugging Face implementation of ModelScopeT2V is surprisingly capable of interpreting prompts and generating content that would likely breach current Chinese regulations on generative models.

stock footage in the training data. The choice to edit hundreds of 4-second clips side by side reflected several aspects of the process: the paired nature of labeled video and image datasets, the dual structure of encoder/decoder architectures, as well as my own experience of comparing outputs of the same prompt. The synthetic video clips are laid over a backdrop of stock footage clips from the datasets used to train ModelScopeT2V, and the resulting collage is accompanied by a narrative script I authored, read by a text-to-speech model trained on my own voice, as well as by a soundtrack mix of Chinese music generated by a text-to-audio model. All of these aspects of *Latent China* would deserve a discussion of their own, but for the purpose of this article I focus on the video clips generated with ModelScopeT2V, reflecting on my creative process from my initial encounter with the generative model, through various experiments with its interface, to my formulation of synthetic probes, from which I will generalize some suggestions for a qualitative approach to the latent spaces of generative models.

The idea of synthetic probes is inspired by methodological discussions across different fields: the development of cultural probes in participatory design, their adoption in Human-Computer Interaction (HCI), the ethnographic use of interview probes, as well as the computer science application of linear classifiers to probe neural networks. In the late 1990s, Gaver et al. (1999) proposed the use of packages of materials, objects and tools, which they called "cultural probes", for participatory research. After being "launched" into a social setting shared by researchers and participants, cultural probes are designed to provoke "a more impressionistic account of their beliefs and desires, their aesthetic preferences and cultural concerns" (p. 25). Two key elements of cultural probes are their development through dialogue between designers and community members (Hemmings et al., 2002) and their embrace of openness and ambiguity (Gaver et al., 2004, p. 56). Probes have been widely adopted and adapted as a research method across HCI research, where they have precipitated substantial debates about the discipline's epistemological commitments (Boehner et al., 2007). The term probe is also used in ethnographic research where it indicates verbal, material or practical prompts designed to "stimulate or encourage an informant to provide data on specific topics with minimal influence from the interviewer" (De Leon & Cohen, 2005, p. 200). As Robert Willim (2017) notes, probes can also bridge between artistic practice and ethnographic research:

> If they should be compared with natural scientific probes, they would have more in common with the kinds that are sent out in the unknown (like space probes), than the ones that are inserted in bodies or objects to precisely capture specimens, samples, or data (p. 213).

Conversely, for computer scientists, probes are closer to measuring instruments with their own trainable parameters that can be used to map what is happening inside machine learning models without influencing their operation (Alain & Bengio, 2018).

In a recent editorial on anthropology and generative models, Anders Kristian Munk notes how the proliferation of algorithmic systems and automated agents has dramatically expanded the scope of ethnographic research, as "a new field has suddenly come into being with its own cultural expressions, its own species of interlocutors, and its own peculiar conditions for doing fieldwork" (2023). A field site might now include computer science labs and user communities, data center rooms and transnational cable networks, machine learning repositories and the latent spaces of generative models, which all have different limitations to access and observation. For example, in contrast to both physical or virtual spaces, a high-dimensional and nonlinear latent space of mathematical vectors exceeds human perception and is not amenable to direct experience (MacKenzie & Munster, 2019), requiring new epistemological and paradigms that can offset its uncertainty and potentiality (Veel, 2021). A wide range of methodological proposals offer different options: reverse-engineering algorithmic black boxes through their inputs and outputs (Diakopoulos, 2015) or by "shaking" them for clues about their operation (Ziewitz, 2016); repurposing their interfaces as research tools (Marres & Gerlitz, 2016); analyzing AI-generated images (Salvaggio, 2023) to unmake the processes behind them (Munn et al., 2023), or talking to large language models to figure out their hermeneutic (Henrickson & Meroño-Peñuela, 2023) or narrative capabilities (Munn & Henrickson, 2024). With my coauthors, we have proposed "synthetic ethnography" (de Seta et al., 2023), a methodological toolbox for the qualitative study of generative models, including "field devices" such as participatory content creation, trace archives, and latent space walks. This essay adds one more field device to this toolbox: the synthetic probe, a purposefully designed object that can be "launched" into a model's latent space to provoke the generation of outputs which can be analyzed iteratively and comparatively (Figure 4).

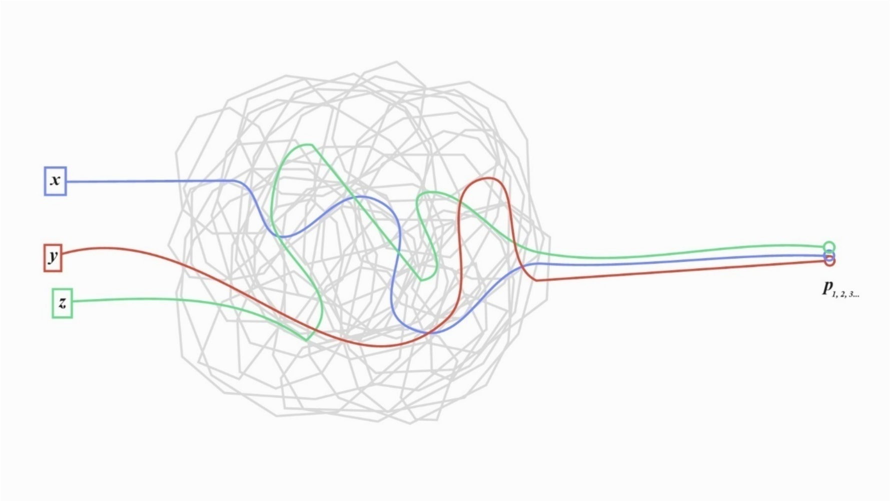

Figure 4. Diagram of synthetic probes (*p*~1, 2, 3~) being launched alongside slightly different trajectories into a machine learning model's high-dimensional latent space (central manifold), resulting in outputs (*x*, *y*, *z*) which can be analyzed comparatively and iteratively to speculate about how the probes' trajectories reflect the model's functioning.

Much like cultural probes, synthetic ones require a dialogic process of design — they are not simple lists of prompts or benchmark tasks to be tested across models or systems. In a similar way to the Twitter bots deployed by Wilkie et al. (2015), synthetic probes are "speculative devices" allowing the sort of interaction which "stimulates latent social realities, and thus facilitate the emergence of different questions" (p. 82). Rather than being developed in collaboration with a community of research participants, synthetic probes are designed through interactions with an algorithmic system; in the case of *Latent China*, they were shaped by the interaction between different elements of ModelScopeT2V (English-language prompting box, CLIP embeddings, LAION and WebVid datasets), by the creative use of the model by early adopters, as well as my own experimentation with it. And just like ethnographic interview probes, they served to stimulate the model in providing data or information about itself. The eighty-plus textual prompts I ended up using for this work developed organically from my intention to explore the China-related area of ModelScopeT2V's latent space (by keeping "China" or "Chinese" as a fixed element in most of them) while also nudging the model into outputting synthetic footage that was representative of its training data through minimalist combinations of words that were neither too broad and hence uninterpretable, nor too narrow and hence overdetermined by my authorial decisions[2]. Probe development was also iterative, as I tested new prompts following small variations and semantic adjacency only once I had achieved some degree of visual space saturation in the generated outputs — for example, by testing "Chinese metropolis" or names of specific Chinese cities only after I had generated enough clips of "Chinese city" to have a representative sample. By the end of this process, after months of daily interactions with ModelScopeT2V, I had developed a sort of intuitive, embodied and perhaps even hallucinatory understanding that was perhaps somewhat close to what artist Everest Pipkin (2020) describes after having watched an entire dataset of one million 3-second videos:

> Very slowly, over and over, my body learns the rules and edges of the dataset. I come to understand so much about it; how each source is structured, how the videos are found, the words that are caught in the algorithmic gathering.

The design of synthetic probes is closely connected to this sort of algorithmic gathering. On the one hand, I converged on prompts that maximized the output of relevant videos and minimized the generation of either noisy or abstract clips. On the other hand, the balance I tried to strike was constantly unsettled by the quirks of the training data, the limitations of the prompting interface, the ambiguity of natural language processing, and the stochasticity of the outputs.

## 5 From Prompting to Probing

In this essay, I have outlined a methodological approach for the qualitative study of generative artificial intelligence models. After introducing the epistemological challenges faced by users of machine learning models, I argued that these black boxed systems can be explored through indirect ways of knowing — or at least guessing — what goes on inside them. Inspired by both ethnographic and digital methods, I proposed the use of what I call *synthetic probes*: qualitative research devices designed to correlate the inputs and outputs of generative models and thus gather insights into their training data, informational representation, and capability for synthesis. In order to ground this proposal in empirical work, I first described the sociotechnical context of a specific text-to-video generative model (ModelScopeT2V), and then explained how my encounter with it resulted in an extensive period of experimentation dedicated to the production of a documentary entirely composed of synthetic video clips. Lastly, reflecting on how this experience bridges between qualitative research and creative practice, I have extrapolated more general observations about how the extensive history of research probes across disciplines can inspire the creation of methodological devices designed to allow the indirect exploration of a machine learning model's latent space. Much like design probes and ethnographic interview probes, synthetic probes are dialogic and open-ended: their trajectory through a model's latent space is meant to precipitate observational data that can shape the refinement of other probes or ground further analyses. At the same time, while sharing with computer science probes the goal of measuring what happens inside the black box of a machine learning model without influencing its operation, synthetic probes are also imprecise instruments. Just as other speculative

[2] Colombo et al. (2023) have proposed a similar approach which they term "ambiguous prompting", which runs counter the common objectives of prompt engineering (generating something as close as possible to the desired outcome) and also reflects a key characteristic of probes: ambiguity.

research methodologies, they are performative rather than descriptive, and might function in unpredictable ways; they are

> designed to "prompt" (as much as probe) emergent enactments that can problematize existing practices […] and open up the prospective. […] They can be grossly alienating as well as playfully confusing, or obliquely inviting: they can, in other words, just as easily precipitate a flight into "the plausible and the probable" by the actors who are being speculatively engaged. There is, therefore, no guarantee that speculative devices and their provocations will work — experiments can and do fail (Wilkie et al., 2015, pp. 98–99).

How can this experiment be replicated in other sociotechnical contexts and algorithmic gatherings? Different assemblages of machine learning models, interfaces, datasets or systems require different probe designs, launch trajectories, and retrieval protocols. Synthetic probes are not necessarily textual prompts: they can include visual content like images or videos, musical snippets or vocal instructions, structured tasks or pieces of code — whatever input is capable of provoking and stimulating the automated agent to output some form of information about its own architecture, functioning, limits, and so on. Probes can be designed to leave traces or return data that can be analyzed comparatively or across synchronic or diachronic axes; they can exploit repeatability (for example, by keeping a fixed seed or parameter) or embrace failure (by pursuing overfitting or hallucinations). They can be developed to compare multiple models or systems, to create feedback loops between them, to exploit their self-referential capabilities, or to reveal their limitations and boundaries. They can be modulated to find out the minimal requirements of input or pushed towards system failure. Most importantly, synthetic probes should not be rationalized as an objective method of inquiry: as Gaver et al. (2004) observed, "we value the mysterious and elusive qualities of the uncommented returns themselves. Far from revealing an 'objective' view of the situation, the Probes dramatize the difficulties of communicating with strangers" (p. 55). As stranger and stranger entities like automated agents and algorithmic systems multiply these communicational difficulties, launching probes into the newly unfolding spaces of data, computation and cognition can perhaps help us open up new trajectories for inquiry. And this dramatization can become an artistic probe in itself — in the case of *Latent China*, by inviting viewers to interpret its assemblage of synthetic images on their own terms.

## References

Alain, G., & Bengio, Y. (2018). Understanding Intermediate Layers Using Linear Classifier Probes (arXiv:1610.01644). *arXiv*. http://arxiv.org/abs/1610.01644

Alibaba Cloud Community. (2022). Alibaba Cloud Launches ModelScope Platform and New Solutions to Lower the Threshold for Materializing Business Innovation. *Alibaba Cloud*. https://www.alibabacloud.com/blog/alibaba-cloud-launches-modelscope-platform-and-new-solutions-to-lower-the-threshold-for-materializing-business-innovation_599467

Barr, K. (2023). Text to Video Generative AI is Finally Here and It's Weird as Hell. *Gizmodo*. https://gizmodo.com/text-to-video-ai-art-generator-runway-modelscope-ai-1850249431

Boehner, K., Vertesi, J., Sengers, P., & Dourish, P. (2007). How HCI Interprets the Probes. In M.B. Rosson & D.J. Gilmore (Eds.), *Proceedings of the SIGCHI Conference on Human*

*Factors in Computing Systems* (pp. 1077–1086). New York, NY: ACM Press. https://doi.org/10.1145/1240624.1240789

chaindrop. (2023). Will Smith Eating Spaghetti [Reddit Post]. R/StableDiffusion. https://www.reddit.com/r/StableDiffusion/comments/1244h2c/will_smith_eating_spaghetti/

Cole, S. (2023). AI Will Smith Eating Spaghetti Will Haunt You For the Rest of Your Life. *Vice*. https://www.vice.com/en/article/xgw8ek/ai-will-smith-eating-spaghetti-hill-haunt-you-for-the-rest-of-your-life

Colombo, G., Niederer, S., de Gaetano, C., & Borie, M. (2023). Prompting Generative Visual AI for Biodiversity: From Prompt Engineering to Prompt Design. Generative Methods – AI as Collaborator and Companion in the Social Sciences and Humanities. Conference, Aalborg University, December 6–8.

Davison, R.M., Chughtai, H., Nielsen, P., Marabelli, M., Iannacci, F., van Offenbeek, M., Tarafdar, M., Trenz, M., Techatassanasoontorn, A.A., Díaz Andrade, A., & Panteli, N. (2024). The Ethics of Using Generative AI for Qualitative Data Analysis. *Information Systems Journal*, *34*(5), 1433–1439. https://doi.org/10.1111/isj.12504

De Leon, J.P., & Cohen, J.H. (2005). Object and Walking Probes in Ethnographic Interviewing. *Field Methods*, *17*(2), 200–204. https://doi.org/10.1177/1525822X05274733

Denton, E., Hanna, A., Amironesei, R., Smart, A., & Nicole, H. (2021). On the Genealogy of Machine Learning Datasets: A Critical History of ImageNet. *Big Data & Society*, *8*(2), 1–14. https://doi.org/10.1177/20539517211035955

de Seta, G., Pohjonen, M., & Knuutila, A. (2023). Synthetic Ethnography: Field Devices for the Qualitative Study of Generative Models. *SocArXiv*. https://doi.org/10.31235/osf.io/zvew4

Diakopoulos, N. (2015). Algorithmic Accountability: Journalistic Investigation of Computational Power Structures. *Digital Journalism*, *3*(3), 398–415. https://doi.org/10.1080/21670811.2014.976411

Elish, M.C., & boyd, danah. (2018). Situating Methods in the Magic of Big Data and AI. *Communication Monographs*, *85*(1), 57–80. https://doi.org/10.1080/03637751.2017.1375130

Gan, W., Wan, S., & Yu, P.S. (2023). Model-as-a-Service (MaaS): A survey (arXiv:2311.05804). *arXiv*. http://arxiv.org/abs/2311.05804

Gaver, B., Dunne, T., & Pacenti, E. (1999). Cultural Probes. *Interactions*, *6*(1), 21–29. https://doi.org/10.1145/291224.291235

Gaver, W.W., Boucher, A., Pennington, S., & Walker, B. (2004). Cultural Probes and the Value of Uncertainty. *Interactions*, *11*(5), 53–56. https://doi.org/10.1145/1015530.1015555

Hemmings, T., Crabtree, A., Rodden, T., Clarke, K., & Rouncefield, M. (2002). Probing the Probes. In T. Binder, J. Gregory & I. Wagner (Eds.), *Proceedings of the Participatory Design Conference* (pp. 42–50). Palo Alto, CA: CPSR.

Henrickson, L., & Meroño-Peñuela, A. (2023). Prompting Meaning: A Hermeneutic Approach to Optimising Prompt Engineering with ChatGPT. *AI & Society*. https://doi.org/10.1007/s00146-023-01752-8

Hoover, A. (2023). AI Videos Are Freaky and Weird Now. But where Are They Headed? *WIRED*. https://www.wired.com/story/text-to-video-ai-generators-filmmaking-hollywood/

Institute for Intelligent Computing. (2023). *[Text-to-video Synthesis Model—English—Public domain]*. ModelScope. https://www.modelscope.cn/models/iic/text-to-video-synthesis

Jiang, J.A., Wade, K., Fiesler, C., & Brubaker, J.R. (2021). Supporting Serendipity: Opportunities and Challenges for Human-AI Collaboration in Qualitative Analysis. In J. Grudin & J. Carroll (Eds.), *Proceedings of the ACM on Human-Computer Interaction* (pp. 1–23). New York, NY: ACM Press. https://doi.org/10.1145/3449168

MacKenzie, A., & Munster, A. (2019). Platform Seeing: Image Ensembles and Their Invisualities. *Theory, Culture & Society*, *36*(5), 3–22. https://doi.org/10.1177/0263276419847508

Marres, N., & Gerlitz, C. (2016). Interface Methods: Renegotiating Relations between Digital Social Research, STS and Sociology. *The Sociological Review*, *64*(1), 21–46. https://doi.org/10.1111/1467-954X.12314

Mok, A. (2023). I Can't Stop Watching These Hilariously Bad AI-Generated Videos of Celebrities Like Will Smith and Scarlett Johansson. *Business Insider*. https://www.businessinsider.com/watch-hilariously-bad-ai-modelscope-videos-will-smith-scarlett-johansson-2023-3

Munk, A.K. (2023). Coming of Age in Stable Diffusion. *Anthropology News*. https://www.anthropology-news.org/articles/coming-of-age-in-stable-diffusion/

Munn, L., & Henrickson, L. (2024). Tell Me a Story: A Framework for Critically Investigating AI Language Models. *Learning, Media and Technology*, 1–17. https://doi.org/10.1080/17439884.2024.2327024

Munn, L., Magee, L., & Arora, V. (2023). Unmaking AI Imagemaking: A Methodological Toolkit for Critical Investigation (arXiv:2307.09753). *arXiv*. http://arxiv.org/abs/2307.09753

Offert, F. (2023). On the Concept of History (in Foundation Models). *IMAGE*, *37*(1), 121–134. https://doi.org/10.1453/1614-0885-1-2023-15462

Pipkin, E. (2020). On Lacework: Watching an Entire Machine-Learning Dataset. Unthinking Photography. https://unthinking.photography/articles/on-lacework

Rogers, R. (2013). *Digital Methods*. Cambridge, MA: MIT Press.

Salvaggio, E. (2023). How to Read an AI Image: Toward a Media Studies Methodology for The Analysis of Synthetic Images. *IMAGE*, *37*(1), 83–89. https://doi.org/10.1453/1614-0885-1-2023-15456

TechNode Feed. (2023). Alibaba's ModelScope Attracts Over 2 Million Developers Amid AI Frenzy. *TechNode*. http://technode.com/2023/08/01/alibabas-modelscope-attracts-over-2-million-developers-amid-ai-frenzy/

Veel, K. (2021). Latency. In N.B. Thylstrup, D. Agostinho, A. Ring, C. D'Ignazio, & K. Veel (Eds.), *Uncertain Archives: Critical Keywords for Big Data* (pp. 313–319). Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/12236.003.0034

Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., & Zhang, S. (2023). ModelScope Text-to-Video Technical Report (arXiv:2308.06571). *arXiv*. http://arxiv.org/abs/2308.06571

Wang, S.C., Van Durme, B., Eisner, J., & Kedzie, C. (2024). Do Androids Know They're Only Dreaming of Electric Sheep? (arXiv:2312.17249). *arXiv*. http://arxiv.org/abs/2312.17249

Wilkie, A., Michael, M., & Plummer-Fernandez, M. (2015). Speculative Method and Twitter: Bots, Energy and Three Conceptual Characters. *The Sociological Review*, *63*(1), 79–101. https://doi.org/10.1111/1467-954X.12168

Will Smith [@WillSmith2real]. (2024). This Is Getting Out of Hand! [Tweet]. *X* (formerly Twitter). https://twitter.com/WillSmith2real/status/1759703359727300880

Willim, R. (2017). Evoking Imaginaries: Art Probing, Ethnography and More-than-Academic Practice. *Sociological Research Online*, *22*(4), 208–231. https://doi.org/10.1177/1360780417726733

Yu, I. (2023). Q&A: Alibaba Cloud's CTO on Creating China's Biggest AI Model Community. *Alizila*. https://www.alizila.com/alibaba-cloud-cto-creating-china-biggest-ai-model-community-llm/

Ziewitz, M. (2016). Governing Algorithms: Myth, Mess, and Methods. *Science, Technology, & Human Values*, *41*(1), 3–16. https://doi.org/10.1177/0162243915608948

**Gabriele de Seta** – Department of Linguistic, Literary and Aesthetic Studies, University of Bergen (Norway)
https://orcid.org/0000-0003-0497-2811 | ✉ gabriele.seta@uib.no
http://paranom.asia

Gabriele de Seta is, technically, a sociologist. He is a Researcher at the University of Bergen (Norway), where he leads the ALGOFOLK project ("Algorithmic folklore: The mutual shaping of vernacular creativity and automation") funded by a Trond Mohn Foundation Starting Grant (2024–2028). His research work, grounded on qualitative and ethnographic methods, focuses on digital media practices, sociotechnical infrastructures, and vernacular creativity.