

# MAP-Elites with Transverse Assessment for Multimodal Problems in Creative Domains

Marvin Zammit<sup>[0000-0002-5343-0077]</sup>, Antonios Liapis<sup>[0000-0001-5554-1961]</sup>, and  
Georgios N. Yannakakis<sup>[0000-0001-7793-1450]</sup>

Institute of Digital Games, University of Malta, Msida, Malta  
{marvin.zammit, antonios.liapis, georgios.yannakakis}@um.edu.mt

**Abstract.** The recent advances in language-based generative models have paved the way for the orchestration of multiple generators of different artefact types (text, image, audio, etc.) into one system. Presently, many open-source pre-trained models combine text with other modalities, thus enabling shared vector embeddings to be compared across different generators. Within this context we propose a novel approach to handle multimodal creative tasks using Quality Diversity evolution. Our contribution is a variation of the MAP-Elites algorithm, MAP-Elites with Transverse Assessment (MEliTA), which is tailored for multimodal creative tasks and leverages deep learned models that assess coherence across modalities. MEliTA decouples the artefacts’ modalities and promotes cross-pollination between elites. As a test bed for this algorithm, we generate text descriptions and cover images for a hypothetical video game and assign each artefact a unique modality-specific behavioural characteristic. Results indicate that MEliTA can improve text-to-image mappings within the solution space, compared to a baseline MAP-Elites algorithm that strictly treats each image-text pair as one solution. Our approach represents a significant step forward in multimodal bottom-up orchestration and lays the groundwork for more complex systems coordinating multimodal creative agents in the future.

**Keywords:** MAP-Elites · Quality Diversity · Image Generation · Text Generation · Text-to-image Generation · Digital Games.

## 1 Introduction

Evolutionary search in creative domains such as visual, audio, or text generation has traditionally struggled to evaluate the artefacts it produces. This is mostly because there is no universal metric to assess the quality of media content [38, 13]. Early approaches relied on ad-hoc metrics such as timing intervals in music generation [1] and compression-based indices for image generation [29], or tasked humans to evaluate the evolving population [41, 27, 45, 20]. As more refined deep learning algorithms became available, models trained for a specific task have been employed more frequently as a fitness measure of evolved artefacts [21, 40].

For generative media, the most interesting development in the field of deep learning is the training and release of multimodal models. These models map

multiple modalities to the same latent space, thereby enabling the direct comparison of different types of media. Contrastive Language-Image Pre-Training (CLIP) [35] was one such model which demonstrated excellent zero-shot image classification to any input set of semantic labels. Similar models map text with other modalities, such as audio [9], which in unison with CLIP (or similar models) may compare images to another modality via an intermediary text modality. Alternatively, models such as Meta’s ImageBind [14] directly combine multiple input and output modalities into a single embedding space, which facilitates *multimodal generation* and assessment but also opens up new possibilities for cross-modal learning and transfer learning.

The advent of more nuanced metrics based on large-scale corpora (unimodal or multimodal) does not quite address the limitations of optimising a universal “quality” metric in creative domains. Such a singular drive may lead to a narrow view of human creativity which often builds on niches such as art movements, music genres, and literary paradigms. To address this, more recent research in evolutionary computation has focused on the diversity of the output instead of its quality [26], or combining the two in Quality Diversity (QD) algorithms [34, 15]. QD algorithms promote diversity in the artefacts while maintaining some minimal criteria on quality [25, 28] or keeping only the fittest individuals within each phenotypic niche [32]. The latter approach is followed in Novelty Search with Local Competition [26], which only compares neighbouring individuals to assess their (local) dominance. Another prominent QD algorithm is Multi-dimensional Archive of Phenotypic Elites (MAP-Elites) [32]. MAP-Elites partitions the solution space into a multi-dimensional grid (the *feature map*), where each axis represents varying properties within a specific behavioural characteristic (BC) or phenotypic trait of the solutions. Each cell stores the optimal individual (elite) according to the global fitness function, promoting only competition within the phenotypic niche. The most popular implementation of MAP-Elites operates in a steady-state fashion, selecting a parent among the elites (at random) and mutating it to produce an offspring. The offspring is then mapped to a cell of the feature map according to its BCs and may replace the elite in that cell if it has a higher fitness. As MAP-Elites illuminates a problem space, it is particularly apt for creative domains where it has already shown successes [12, 15, 49, 2, 3, 8].

This paper applies the MAP-Elites algorithm to a multimodal creative domain, specifically generating text descriptions and cover images for hypothetical video games. To address this challenge, we propose an algorithmic improvement on QD search: MAP-Elites with Transverse Assessment (MEliTA). MEliTA introduces an inter-modal evaluation process that shares partial artefacts (e.g. image or text) among phenotypically similar elites in order to find more coherent pairings. This innovative approach enhances the creative co-evolutionary process, resulting in the discovery of fitter and more diverse outcomes.

## 2 MAP-Elites with Transverse Assessment

MAP-Elites with Transverse Assessment (MELiTA) is a variant of MAP-Elites designed to evolve multimodal artefacts to minimise incongruity between modalities. MELiTA builds on a number of assumptions, which in our use case revolve around two modalities but can scale to any number of modalities:

- each evolved individual ( $A$ ) is a collections of  $N$  (separable) artefacts, each encompassing a single modality  $M_i$ , e.g.  $A = \{a_{M1}, a_{M2}, \dots\}$
- variation operators ( $V_{M1}, V_{M2}, \dots$ ) can be applied on each artefact type ( $a_{M1}, a_{M2}, \dots$ ) separately, potentially informed by other modalities but not modifying the other artefact types, e.g.  $a'_{M1} = V_{M1}(a_{M1}, a_{M2}, \dots)$  etc.
- there exists a function ( $f$ ) that returns a value  $q$  indicating coherence between all modalities, e.g.  $q(A) = f(a_{M1}, a_{M2}, \dots)$
- there exist  $N$  functions ( $g_{M1}, g_{M2}, \dots$ ) each returning a value ( $\beta_{M1}, \beta_{M2}, \dots$ ) indicating properties of one artefact type ( $a_{M1}, a_{M2}, \dots$ ) separately, e.g.  $\beta_{M1}(A) = g_{M1}(a_{M1})$ ,  $\beta_{M2}(A) = g_{M2}(a_{M2})$ , etc.

Following the above notations, MELiTA produces an  $N$ -dimensional archive of elites ( $N$  being the number of modalities in the artefacts), characterised by  $N$  behaviour characterisations ( $\beta_{M1}, \beta_{M2}, \dots$ ).

During the evolutionary cycle, an existing elite  $E = \{e_{M1}, e_{M2}, \dots\}$  is selected from the archive—the selection operator can be uniform selection or more sophisticated [10]. One artefact of this individual is chosen randomly (e.g.  $e_{M1}$ ) and changed via the appropriate variation operator, creating in this example  $e'_{M1} = V_{M1}(e_{M1}, e_{M2}, \dots)$ . The new artefact ( $e'_{M1}$ ) is assigned a behaviour characterisation based on its modality (i.e.  $g_{M1}(e'_{M1})$ ). A new individual  $E'$  is created by combining the new artefact with unchanged artefacts of the parent  $E$ , i.e.  $E' = \{e'_{M1}, e_{M2}, \dots\}$ . In vanilla MAP-Elites (as applied in this paper), the individual  $E'$  would be compared with the elite ( $E'_{old}$ ) with BCs  $g_{M1}(e'_{M1}), g_{M2}(e_{M2}), \dots$  in terms of  $q$ , replacing it if  $q(E') > q(E'_{old})$  or occupying the cell if no elite exists for those BCs. In MELiTA, the artefact  $e'_{M1}$  is iteratively paired with the artefacts of other modalities of each elite  $R = \{r_{M1}, r_{M2}, \dots\}$  that occupies a cell with BC  $g_{M1}(e'_{M1})$ , producing a new candidate solution  $R' = \{e'_{M1}, r_{M2}, \dots\}$  and computing the coherence between modalities for the new individual ( $q(R')$ ). The collection of candidate solutions  $\mathbf{R}'$  along with  $E'$  are sorted by their coherence score  $q$  and form an ordered list of candidate solutions  $\mathbf{L}$ . In order, each member  $\phi \in \mathbf{L}$  is checked against the occupying elite ( $\phi_{old}$ ) in a cell with the BCs of  $\phi$ . If no such elite exists (the cell is empty), the individual  $\phi$  occupies that cell and the process ends. If the current elite  $\phi_{old}$  is worse than the new candidate ( $q(\phi_{old}) < q(\phi)$ ) then  $\phi$  replaces  $\phi_{old}$  and the process ends; if the current elite is not worse, then the process continues with the next member in  $\mathbf{L}$ . This results in only one new individual being inserted into the archive (at most) per evolutionary cycle, and ensures that empty cells can also be filled in the archive (by individual  $E'$ , if it is better than other alternatives).

The process of MELiTA will become clearer through the use case of Section 3.

### 3 Use Case: Generating Text & Visuals for Game Titles

As an exploratory use case, we select images and text as modalities for MELiTA, as both benefit from the availability of reliable AI generators. In this use case, the goal is to generate fictitious video games in the form of a *game title*, a *short description* of the game, and a *cover image*. Through this experiment, we wish to create a diverse set of coherent and appropriately game-like art and text blurbs that can inspire players and game developers alike.

For each modality, core considerations are how the artefact is generated (or changed via mutation) and how it is characterised for the purposes of the MAP-Elites feature map [32]. The sections below clarify how artefacts of each modality are generated and characterised, followed by a rundown of the MELiTA process.

#### 3.1 Text Modality

The Generative Pre-trained Transformer 2 (GPT-2) [36] is an auto-regressive model based on the transformer architecture [48]. This model has undergone extensive pre-training through a substantial corpus of English text through a self-supervised learning approach [4]. While GPT-2 has since been eclipsed by more cogent models [6], particularly the more recent large-language models (LLMs) [33, 46, 47], it distinguishes itself with considerably faster inference, at the expense of reduced performance. For an iterative evolutionary algorithm, the substantial speed gain GPT-2 offers was considered a good trade-off.

**Text Generation.** To generate believable titles and descriptions for fictitious games, a GPT-2 language model was fine-tuned on a dataset composed of real titles and descriptions from the extensive catalogue of the Steam platform<sup>1</sup>. The data was curated, removing entries without English text, non-game entries (e.g. utilities, videos), and entries labelled with adult material or nudity. The resulting dataset contains approximately 72,000 pairs of game titles and descriptions.

For game titles, the pre-trained GPT-2 model of [36] was fine-tuned exclusively on video game titles of the above Steam dataset. Given the modest scale of this dataset, we used the most compact variant of the model (approximately 124 million parameters). The transformer was trained on the list of titles demarcated by distinctive beginning and end tokens: “<|begin|>game title<|end|>”. First attempts exhibited over-fitting, and the model echoed existing titles. To mitigate this, the weights and biases of the last layer were reset prior to the fine-tuning process. This resulted in a more robust model capable of generating novel titles exhibiting minimal overlap with their original counterparts.

A second GPT-2 model was fine-tuned from the pre-trained model of [36], this time incorporating both title and description (see example in Table 1) in the format “<|begin|>game title<|body|>description<|end|>”. A new game title can thus be generated through the first model using an input prompt of “<|begin|>”. A description can be generated by priming the second model with this new title in the format “<|begin|>game title<|body|>”.

<sup>1</sup> <https://store.steampowered.com/>

Table 1: Text variation samples: Partial mutation chooses a space or punctuation at random, removes the text after it and uses the sequence up to that point to generate the remaining description (in red). Full mutation removes the description and only uses the game title (in bold) to generate the description (in red).

Original	<b>Hooley! You Got a Monster!?</b> , You've been selected for an experiment at the University of Chicago's Animal Research Center. A girl, who has lost her memory in this creepy and magical place will use your memories to discover why you got kidnapped by Professor Teller on his way there...
Partial mutation	<b>Hooley! You Got a Monster!?</b> , You've been selected for an experiment at the University of Chicago's Animal Research Center. <b>A professor with no memories about his past finds you. An unforgettable story about two animals, their lives and what happened to them in this dream-like place.</b>
Full mutation	<b>Hooley! You Got a Monster!?</b> , <b>The best game of all time is back, and better than ever. This year we've reimaged the classic arcade platformer with over 100 levels to conquer. Gameplay has been revamped for even more fun, it's more difficult then ever but much harder in this new version!</b>

**Text Mutation.** Two variation operators may be applied during mutation of text descriptions: one (partial mutation) retains coherence with the previous description while the other (full mutation) resets the description in order to avoid early convergence. For partial mutation, the first part of the description was retained, split along a selected space or punctuation mark around the middle of the text<sup>2</sup>. The rest of the description is removed. The game title together with the first description fragment are reintroduced as input to the second GPT-2 model to complete the description. For full mutation, the entire description is removed and only the game title is used as input to the second GPT-2 model to produce a new description from scratch. The chance of full mutation is 20%, otherwise partial mutation is applied. Text variation samples are shown in Table 1.

**Text Characterisation.** Several approaches to characterising text were explored for this use case, e.g. via the Gunning-Fog readability index [16]. We settled on topic modelling using topics extracted from the Steam dataset, as we want to match against dominant patterns for game descriptions, typical [38] of the domain we attempt to emulate. We used the Latent Dirichlet Allocation [5] algorithm (LDA) for topic modelling and trained the algorithm across different numbers of topics (from 4 to 30). Using the perplexity and complexity metrics as guides, we settled on an optimal setup of 16 topics.

To characterise each generated description, we subject it to the tuned LDA topic model, which results in a set of probabilities designating the likelihood of the description to be aligned with each of the predefined topics. In situations where the most probable topic assignment falls short of achieving a probability threshold of at least 40% above other topics, the description is deemed as *unclassified* and is not added to the feature map (i.e. it is ignored during evolution).

### 3.2 Image Modality

We leverage Stable Diffusion (SD) for the image generation tasks in this paper as it can produce high-quality images with low compute. SD [39] has openly

<sup>2</sup> We randomly select three spaces or punctuation marks within the text and keep the middle one. This makes it likely that the split will be in the middle of the description.

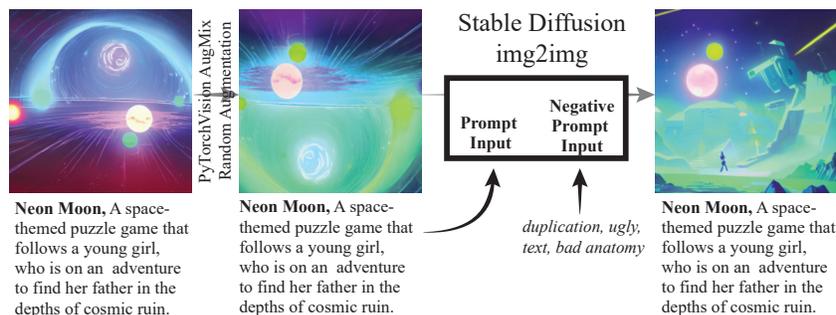


Fig. 1: Image variation sample: The parent’s (unchanged) text modality is used as a prompt for image repair based on SD, alongside “standard” negative prompts.

accessible source code and pre-trained weights. Where diffusion models [44] are trained to reconstruct noisy versions of the desired output (e.g. image), SD models are trained on a noisy latent vector of the output, thus reducing inference time and improving robustness. Using a classifier-free guidance approach [19], both a conditional and an unconditional diffusion model are trained, and by comparing their responses during inference, a balance can be struck between image quality and faithful image adherence to the input prompt.

**Image Generation.** For this paper we utilise a text-to-image SD model to generate the cover art for the fictitious games. The input textual prompts contain both a hypothetical game title and a corresponding description, presented in the format “**title, description**” (see example in Table 1). Based on early experiments, we also add a set of negative prompts (“duplication, ugly, text, bad anatomy”) to enhance the aesthetic quality of the generated outputs.

**Image Mutation.** As with text, we mutate the image on the phenotype level in two stages. First, we apply the AugMix augmentation function [18] from the TorchVision software library [30] to distort the original image. The distorted image is then paired with the original prompts (game title, description, negative prompts) as inputs to an image-to-image SD model (see Fig. 1). In order to strike a balance between image quality and computational efficiency, all tasks pertaining to image generation and mutation were executed over a sequence of 40 diffusion steps. Since the augmentation function may lead to more or less distorted images, resulting images after SD may match the parent image to a lesser or greater degree. Unlike text mutation, we have less control over the chances of a large phenotypic change, but consider such a change beneficial to avoid early convergence and a slow evolutionary process.

**Image characterisation.** As with text classification we explored several BCs for images, but we settled on two straightforward metrics: complexity and colour-

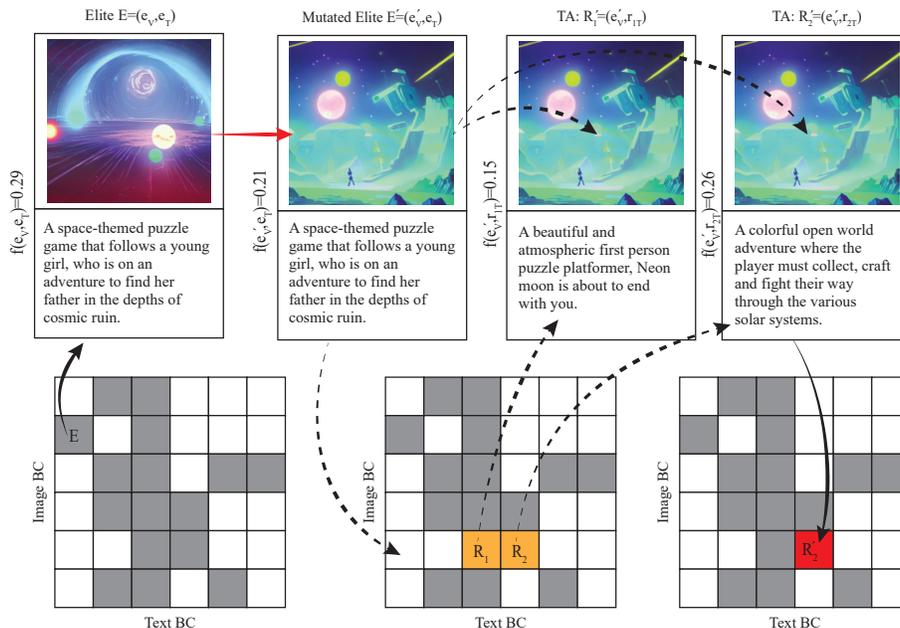


Fig. 2: The MELiTA process in a simplified feature map for this use case, with grey cells occupied by elites. From one selected elite  $E$ , the changed image ( $e'_v$ ) produces three candidate solutions from elites  $E$ ,  $R_1$ ,  $R_2$ . Based on their CLIP score, the ordered list of candidates is  $L = \{R'_2, E', R'_1\}$ . Since  $q(R'_2) > q(R_2)$  the candidate  $R'_2$  (that merges the image from  $E'$  and text from  $R_2$ ) replaces  $R_2$ . If  $q(R'_2) \leq q(R_2)$  then  $E'$  would occupy the empty cell at  $(5,0)$ . Dotted lines denote temporary individuals that are lost after this parent selection.

fulness. *Image complexity* is calculated as the ratio of edge pixels found via the Holistically-Nested Edge Detection (HED) model [51] over the total pixel count. While previous approaches in evolutionary art relied on Sobel or Canny filters [29] for complexity estimation, HED edges offer a notable advantage in terms of accuracy and noise reduction. *Image colourfulness* is calculated via a quantitative measure of the perceived chromatic richness or saturation [17]. To combine the two image metrics into a concise BC, each numerical value (complexity, colourfulness) is categorised into four bins (very low, low, high, very high) and the image is classified in terms of the combinations of these bins. Therefore, the image BC comprises a total of 16 bins, matching the dimensions of the text BC.

### 3.3 MELiTA applied to the use case

MELiTA in this bimodal use case performs parent selection via an Upper Confidence Bound (UCB) algorithm [23], which takes into account frequency of parent selection and can improve coverage of the final elites [42]. The feature map con-

sists of a grid of  $16 \times 16$  cells, with a text BC for membership in one of 16 topics found in the Steam database (see Section 3.1) and a visual BC for membership in one of 16 combinations of image complexity and colourfulness categories (see Section 3.2). This use-case uses the CLIP score between image and description (including the title as shown in Table 1) as fitness, to decide whether elites survive in the archive. In ‘vanilla’ MAP-Elites the selected parent produces either a new image or a new description via variation operators, then pairs it with the remaining artefact (description or image, respectively) in its genotype to produce a new offspring. However, Transverse Assessment operates as follows.

When a parent is selected in MELiTA, either a new image or a new description (chosen randomly) is produced from the parent’s existing artefacts via variation operators described in Sections 3.1 and 3.2. Figure 2 shows an example where the image ( $e_V$ ) is modified (to  $e'_V$ ), and will be described below. The new image  $e'_V$  is paired with the parent’s existing text ( $e_T$ ) and the CLIP score, text BC and visual BC coordinates of this new candidate solution  $E'$  are calculated (in this case, placing it at the currently empty cell at 5,0). For transverse assessment the image is also paired with the title and description of all elites with the same image BC of  $E'$  (the row with  $R_1, R_2$  in Fig. 2), creating new candidate solutions; their CLIP score is calculated<sup>3</sup>. Starting from the fittest of these temporary solutions, if the CLIP score of an existing elite in the archive with these coordinates is lower than the new solution or if the cell is unoccupied, the candidate solution occupies that cell and the process stops (in this case replacing  $R_2$ ). If no candidate solution is added to the archive, the process ends and a new parent selection is made.

## 4 Experimental Protocol

In order to evaluate the performance of MELiTA as a QD algorithm, we leverage the bimodal generation challenge described in Section 3 and aim to validate the high-level hypothesis that *MAP-Elites with Transverse Assessment leads to better quality and diversity in the generated artefacts than MAP-Elites without Transverse Assessment*. We formalise this into more concise hypotheses:

- H1 The final archive of solutions is better and more diverse for MELiTA compared to MAP-Elites without transverse assessment.
- H2 MELiTA discovers better and more diverse solution faster than MAP-Elites without transverse assessment.
- H3 The final archive of solutions for MELiTA can be perceived as more diverse by humans compared to MAP-Elites without transverse assessment.

H1 assesses the final product of evolution, while H2 assesses the performance of the evolutionary process over time. Both H1 and H2 assess diversity based on the ad-hoc BCs of the feature map, while H3 uses orthogonal diversity metrics which are aligned to human perceptions of diversity per modality.

<sup>3</sup> The BC coordinates for these candidate solutions do not need to be recalculated as they are combinations of text BCs and visual BCs that are already known.

**Performance Metrics.** To evaluate H1 and H2, we rely on traditional performance metrics for QD evolutionary search [32]. Specifically, for H1, we evaluate the final archive’s mean fitness (among occupying elites), max fitness (i.e. the highest fitness among elites), coverage (i.e. ratio of occupied cells over all cells in the grid) and QD score (i.e. sum of all the elites’ fitness). The first two metrics measure quality, coverage measures diversity (according to the chosen BCs) and QD score measures a combination of the two<sup>4</sup>. For H2, the same metrics are tracked over time (i.e. after every parent selection) and assessed as an area under the curve (AUC); high AUC values may mean that the final metrics have high scores or that reaching a high score was done earlier in the process. For H3, the final elites are assessed in terms of orthogonal measures of diversity to those used to populate the feature map. For visual diversity two measures are used: the Learned Perceptual Image Patch Similarity (LPIPS) measure is used as a distance metric trained on human annotations of visual distance [53], and the structural similarity (SSIM) measure commonly used for quantifying image degradation suffered through transmission or compression losses. For textual diversity in the descriptions, their embeddings generated by the SBERT encoder [37] are employed, since they capture the semantics of each sentence. Latent vectors of two individuals’ descriptions are compared in terms of cosine similarity [11] to derive the *SBERT distance*. For each distance metric, we calculate for each elite the mean distance with all other elites, as well as the nearest-neighbour distance with the closest elite as a more reserved measure of visual/text overlap.

**Test Cases.** In order to verify these hypotheses, we follow the below protocol. We generated 100 game titles via the GPT-2 model described in Section 3.1 and select 7 titles which are varied in terms of theme and length. These titles are:

- T1 “Neon Moon”
- T2 “Lion King”
- T3 “Hexgrave”
- T4 “Fantasy Fables: The Legend of the Flying Sword”
- T5 “The Princess of Thieves”
- T6 “The Shadow Warrior 2: Shadows of the Past”
- T7 “Hooey! You Got a Monster!?”

For each title, we perform 10 evolutionary runs per tested method (MEliTA and MAP-Elites). To provide a fair and controllable initial population for these methods, we use GPT-2 to produce 100 descriptions per title and for each description we produce 4 images via SD. In each evolutionary run, all text descriptions and a random image among the 4 candidates per description produce the initial 100 individuals which are then assigned to the feature map.

<sup>4</sup> Unlike [32], we do not normalise the values to the maximum found across runs and across methods. Instead, we present the non-normalised results (e.g. the ratio of occupied versus the maximum size of the feature map for coverage).

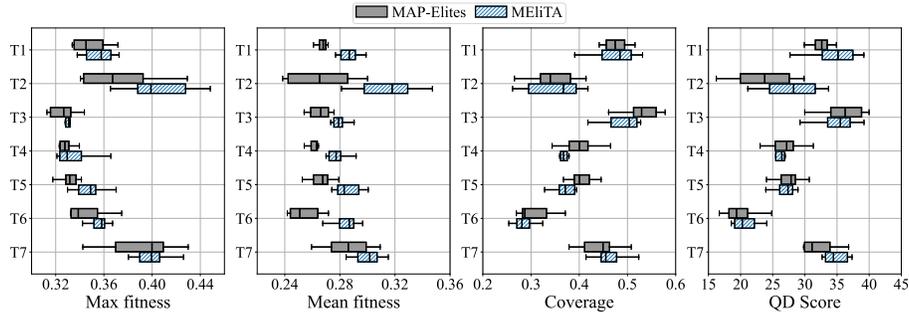


Fig. 3: Metrics of the archives after 2000 selections in MAP-Elites and MELiTA. Box plots summarise values from 10 runs per title.

## 5 Results

In order to validate the hypotheses of Section 4, MELiTA and MAP-Elites with no Transverse Assessment were ran with the same initial populations per game title (minor variation introduced through stochastic image selection as described in Section 4) for 2000 parent selections. Results throughout this section referring to the final archive are derived from the elites in the feature map after 2000 parent selections. Statistical significance between the results of 10 evolutionary runs of different methods is established via the non-parametric two-tail Wilcoxon Rank-Sum test, at a significance level  $p < 0.05$ .

### 5.1 Evaluating the quality and diversity in the final archive

Figure 3 shows the four performance metrics of the QD algorithm applied on the final archive (after 2000 parent selections). It is evident that overall MELiTA results in fitter individuals than MAP-Elites, with results from 6 of 7 game titles having a statistically higher mean fitness (all except T7) and 3 of 7 game titles having a statistically higher maximum fitness (T2, T5, T6). This indicates that the evolved solutions are overall more coherent between the visuals and the text, likely due to the fact that a generated text or image can be paired with another image or text from the archive that is a better fit than the image-text combination produced by MAP-Elites. This prioritisation of pairing artefacts from different elites together leads to a lower coverage of the feature space. While for many game titles this drop in coverage is slight, MAP-Elites has significantly higher coverage for 2 of 7 game titles (T3 and T5). Since MELiTA produces fewer but fitter elites compared to MAP-Elites, QD scores of the two methods tend to be comparable. The only significant difference in QD score is for T7 where MELiTA has a higher QD score than MAP-Elites. Based on this analysis, we can claim that MELiTA leads to better results at the cost, at times, of feature map coverage. Therefore, H1 is only partially validated—for quality but not diversity.

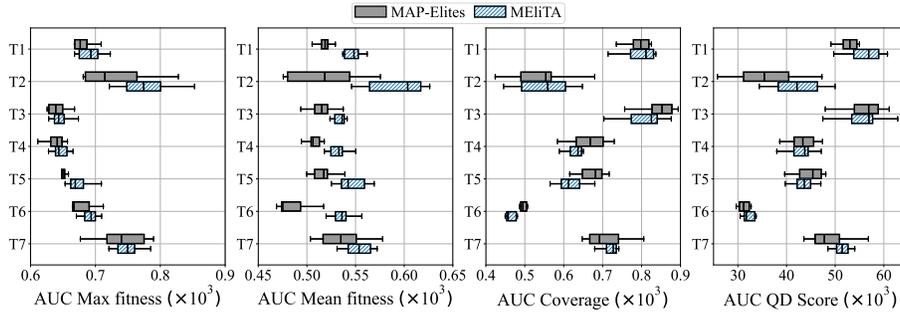


Fig. 4: Area under curve (AUC) of QD metrics over 2000 selections in MAP-Elites and MELiTA. Box plots summarise values from 10 runs per title.

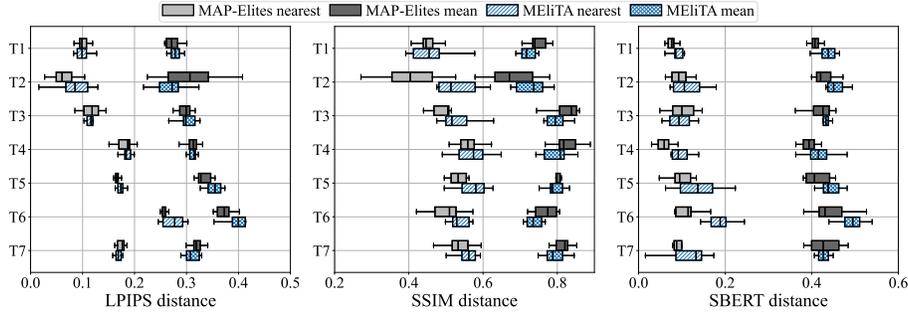


Fig. 5: Visual and textual distance metrics (mean and nearest-neighbour) among final elites of MELiTA and MAP-Elites without Transverse Assessment. Box plots summarise values from 10 runs per title.

## 5.2 Evaluating quality and diversity throughout evolution

Figure 4 shows the area under the curve (AUC) scores over 2000 parent selections for the different QD performance metrics. The findings from the AUC corroborate those of Section 5.1: mean and maximum fitness rises faster in MELiTA than in MAP-Elites (significantly so in 3 and 4 out of 7 game titles respectively). Coverage rises faster in MAP-Elites than MELiTA in most cases (significantly so in 2 out of 7 game titles), and AUC of the QD score is mostly comparable between the two methods; MELiTA has significantly higher AUC for the QD score only for T1. Based on these findings, we can claim that MELiTA can find fitter individuals quicker than MAP-Elites without Transverse Assessment. H2 is thus only partially validated—for quality but not diversity.

## 5.3 Evaluating visual and textual diversity of final artefacts

To assess the perceivable differences in the two methods’ archives, we use visual distances (via LPIPS and SSIM) and textual distance (via SBERT embeddings)

Table 2: Example output of MELiTA for the fictitious game title “The Shadow Warrior 2: Shadows of the Past” (T6). The description (after removal of the title) is left with its original errors.

Game 1	Game 2	Game 3	Game 4	Game 5
				
In this game, players have to use their sword and weapons wisely because there are many enemies who will try very hard for you. After many defeats in a long time, each one offers new challenges.	After a global catastrophe, you are left alone to face your past. The greatest evil - an alien race called Korda - is planning on tearing Earth apart as they do every few centuries in this new adventure from Arkane Studios and Infamous Games.	Play as a shadow warrior, taking on different threats in this third person shooter that's more brutal than ever before.	In this game, players have to use their swords and bows in order not only survive but also fight with others.	After years' long struggle, Dark Lord Arthur's dark legacy is coming to light, and a new evil seems lurking at every corner. TheShadow Warrior2 : Shadows OfthePast.

of the final elites in each run of MELiTA and MAP-Elites after 2000 parent selections. Figure 5 shows the mean and nearest-neighbour distances for each of these metrics, which are largely orthogonal to the BCs used in MAP-Elites.

Figure 5 indicates that, as a whole, the final results of MAP-Elites are not more or less diverse than MELiTA. In terms of visuals, MELiTA has significantly higher mean and nearest-neighbour LPIPS distance than MAP-Elites in results of 2 game titles (T5, T6) and 1 game title (T5) respectively, and significantly higher nearest-neighbour SSIM distance than MAP-Elites in 2 game titles (T2, T3). While these findings are hardly consistent, they indicate that MELiTA tends to produce more diverse images than MAP-Elites. Clearer patterns are gleaned for textual diversity: MELiTA has significantly higher mean and nearest-neighbour SBERT distances in results of 3 game titles (T1, T2, T6) and 4 game titles (T1, T4, T5, T6) respectively. No game title has significant diversity improvements for MELiTA across all metrics. However, we can claim that H3 is validated: the final archives of MELiTA, even if smaller, are more visually and textually diverse than MAP-Elites without Transverse Assessment.

It is also interesting to observe differences in elites directly. We choose T6 (“The Shadow Warrior 2: Shadows of the Past”) as it has high LPIPS and SBERT distances overall, and an indicative run that had high values for both. We apply  $k$ -medoid clustering using a Euclidean distance combining SBERT and LPIPS. The medoids for  $k = 5$  are shown in Table 2. These samples’ text descriptions vary, with some merely describing the gameplay (Games 1, 3 and 4) and some describing a science fiction narrative (Game 2) or a Camelot narrative (Game 5). Some text descriptions include “noise” (e.g. the end of the description in Game 5) while others overlap due to the partial mutation operation (e.g. Game 1 and Game 4). In terms of the images, most depict “action shots” which are likely

not part of gameplay, although Game 5 does show a first-person shooter view (presumably in-game) complete with a gun within an otherwise fantasy setting. Depicted characters are mostly armoured and their faces covered, although it is unclear whether this is because of ninja tropes or because of the negative prompts on e.g. “bad anatomy”. Characters are often seen wielding swords, although there are also modern guns shown in Games 3 and 5. While backgrounds mostly show a foggy forest setting, Game 2 deviates with warmer colours and fiery “wings” on the character at the image centre. Overall, while the images of these medoids do not depict as much diversity as one would expect (in terms of “ninja” characters and colour palettes), there is an overall consistency between the (sometimes generic) descriptions and the associated images. Perhaps the most concerning fact is that Game 2 includes names of actual game studios, likely due to the Steam dataset that GPT-2 was trained on; such text could be problematic due to intellectual property concerns if the use case would be made widely available.

## 6 Discussion

Through the use case for generating fictional game descriptions and cover images, we ascertained that MELiTA can produce fitter—if fewer—elites. We also recognise that coverage decreases since new elites are only added to the archive if there is no better alternative when pairing an offspring’s changed modality with the modalities of an existing elite. This drop in coverage did not result in a drop in visual and textual diversity of produced elites when assessed on metrics decoupled from the BCs. Both in terms of mean distance and nearest-neighbour distance (which can counter the impact of a different number of elites), the elites of MELiTA were more visually and semantically diverse, as well as more coherent (higher fitness) than their counterparts produced without transverse assessment. The diversity metrics (SBERT, LPIPS, SSIM) represent the state-of-the-art for these purposes, but should be corroborated with human feedback. Future work could explore, via a user study with players and game developers, to which degree the generated multimodal artefacts are deemed diverse or inspirational.

It is worth noting that MELiTA produces far more candidate solutions (via transverse assessment) than the single offspring produced by MAP-Elites. This means that for MELiTA the number of fitness evaluations can be much higher than for MAP-Elites, although the number of BC evaluations is the same (only one individual with the parent’s unchanged modality). For a  $16 \times 16$  feature map, as in this use case, MELiTA may perform as many as 16 fitness evaluations per parent selection compared to 1 fitness evaluation for MAP-Elites. Calculating the CLIP score is not expensive, so the computational overhead of MELiTA is negligible; this may not be the case for simulation-based fitness evaluation [32].

The use case in this paper included a crude form of constrained optimisation [7], as the death penalty was applied on all unclassified text descriptions [31]. Additions to MELiTA could explore better ways of handling infeasible individuals, e.g. via a two-population approach as in [43, 22]. Preliminary experiments using additional cells for unclassified individuals did not yield any substantial

differences from MELiTA, but future work could combine constrained QD [43] with a minimal fitness threshold to distinguish feasible individuals.

Examples of the generated results in Table 2 also highlighted some limitations in the chosen variation operators. By operating on the phenotype (images or text) rather than on a latent representation, variation operators produce more controllable and less noisy output—compared to latent variable evolution [12, 52]. Text variation specifically tends to cause overlap between individuals’ descriptions due to partial mutation (see Section 3.1). Alternative text mutation operators could leverage models capable of generating text tokens in reverse [50], i.e. generating the first part of the descriptions given the end part, or use genotypic operators which alter the embeddings of the text [24]. Moreover, the GPT-2 trained model is admittedly dated in the current ecosystem of LLMs. While future work could explore a more modern text generator such as OpenAI’s ChatGPT [33], the main challenge is its closed-source nature which hinders both fine-tuning its weights and keeping track of the provenance of its output.

While this use case explored a bimodal problem that was easy to visualise in a feature map with two axes, MELiTA could conceivably work with more artefact modalities and more BC dimensions—and those two do not need to necessarily match. Each artefact modality could easily have more than one BC; even in our case the two metrics for images (complexity and colourfulness) could have become different BC dimensions leading to a 3-dimensional feature map. Moreover, MELiTA could be applied to a more complex multi-modal creative problem such as evolving text, image, and audio for an interactive story application. In that case, an LLM can be used to generate the story and the locations over which it takes place, SD can be used to generate background images per location, and a text-to-music generator such as MusicGen [9] can generate a soundscape for each part of the story. In this case fitness can be assessed via a multimodal network such as ImageBind [14], by computing the vector similarity of each embedding.

## 7 Conclusion

This paper aimed to address the recent emergence of image-to-text and text-to-image generators through a QD perspective. To this end, we adapt MAP-Elites to operate on solutions that consist of artefacts of different modalities, and implement a transverse assessment method that allows such (partial) artefacts to be shared with other individuals. We show that MELiTA can outperform MAP-Elites, at the cost of fewer solutions. Extensions of this work should explore more state-of-the-art (but open-source) text generation algorithms while also extending the transverse assessment approach to incorporate quality constraints and integrate more types of artefacts and modalities.

## 8 Acknowledgements

This project has received funding from the European Union’s Horizon 2020 programme under grant agreement No 951911.

## References

1. Alfonseca, M., Cebrián, M., De la Puente, A.: A simple genetic algorithm for music generation by means of algorithmic information theory. In: Proceedings of the IEEE Congress on Evolutionary Computation. pp. 3035–3042 (2007). <https://doi.org/10.1109/CEC.2007.4424858>
2. Alvarez, A., Dahlskog, S., Font, J., Togelius, J.: Empowering quality diversity in dungeon design with interactive constrained MAP-Elites. In: Proceedings of the IEEE Conference on Games (2019). <https://doi.org/10.1109/CIG.2019.8848022>
3. Alvarez, A., Font, J.: TropeTwist: Trope-based narrative structure generation. In: Proceedings of the Foundations of Digital Games conference (2022). <https://doi.org/10.1145/3555858.3563271>
4. Balestriero, R., Ibrahim, M., Sobal, V., Morcos, A., Shekhar, S., Goldstein, T., Bordes, F., Bardes, A., Mialon, G., Tian, Y., Schwarzschild, A., Wilson, A.G., Geiping, J., Garrido, Q., Fernandez, P., Bar, A., Pirsiavash, H., LeCun, Y., Goldblum, M.: A cookbook of self-supervised learning. arXiv preprint arXiv:2304.12210 (2023). <https://doi.org/10.48550/arXiv.2304.12210>
5. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. *Journal of machine Learning research* **3**(Jan), 993–1022 (2003)
6. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In: Proceedings of the Neural Information Processing Systems Conference (2020)
7. Coello Coello, C.A.: Constraint-handling techniques used with evolutionary algorithms. In: Proceedings of the Genetic and Evolutionary Computation Conference (2010)
8. Colton, S.: Evolving neural style transfer blends. In: Proceedings of the International Conference on Computational Intelligence in Music, Sound, Art and Design (2021)
9. Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Adi, Y., Défossez, A.: Simple and controllable music generation. arXiv preprint arXiv:2306.05284 (2023)
10. Cully, A., Demiris, Y.: Quality and diversity optimization: A unifying modular framework. *IEEE Transactions on Evolutionary Computation* **22**(2), 245–259 (2017)
11. Dangeti, P.: *Statistics for Machine Learning*. Packt Publishing (2017)
12. Fontaine, M.C., Nikolaidis, S.: Differentiable quality diversity. In: Proceedings of the Neural Information Processing Systems Conference (2021)
13. Galanter, P.: Artificial intelligence and problems in generative art theory. In: Proceedings of the Conference on Electronic Visualisation & the Arts. pp. 112–118 (2019). <https://doi.org/10.14236/ewic/EVA2019.22>
14. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: ImageBind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
15. Gravina, D., Khalifa, A., Liapis, A., Togelius, J., Yannakakis, G.N.: Procedural content generation through quality-diversity. In: Proceedings of the IEEE Conference on Games (2019)

16. Gunning, R.: The technique of clear writing, p. 36–37. McGraw-Hill Book Co. (1973)
17. Hasler, D., Suesstrunk, S.: Measuring colourfulness in natural images. In: Proceedings of the Conference on Electronic Imaging (2003). <https://doi.org/10.1117/12.477378>
18. Hendrycks, D., Mu, N., Cubuk, E.D., Zoph, B., Gilmer, J., Lakshminarayanan, B.: AugMix: A simple data processing method to improve robustness and uncertainty. In: Proceedings of the International Conference on Learning Representations (ICLR) (2020)
19. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: Proceedings of the NeurIPS Workshop on Deep Generative Models and Downstream Applications (2021)
20. Hoover, A.K., Szerlip, P.A., Stanley, K.O.: Interactively evolving harmonies through functional scaffolding. In: Proceedings of the Genetic and evolutionary Computation Conference (2011)
21. Johnson, C.G.: Stepwise evolutionary learning using deep learned guidance functions. In: Proceedings of the International Conference on Innovative Techniques and Applications of Artificial Intelligence. pp. 50–62. Springer International Publishing (2019)
22. Khalifa, A., Lee, S., Nealen, A., Togelius, J.: Talakat: Bullet hell generation through constrained Map-Elites. In: Proceedings of the Genetic and Evolutionary Computation Conference (2018)
23. Kocsis, L., Szepesvári, C.: Bandit based Monte-Carlo planning. In: Proceedings of the European Conference on Machine Learning (2006)
24. Lehman, J., Gordon, J., Jain, S., Ndousse, K., Yeh, C., Stanley, K.O.: Evolution Through Large Models, p. 331–366. Springer Nature Singapore (2023). [https://doi.org/10.1007/978-981-99-3814-8\\_11](https://doi.org/10.1007/978-981-99-3814-8_11)
25. Lehman, J., Stanley, K.O.: Revising the evolutionary computation abstraction: Minimal criteria novelty search. In: Proceedings of the Genetic and Evolutionary Computation Conference (2010)
26. Lehman, J., Stanley, K.O.: Evolving a diversity of virtual creatures through novelty search and local competition. In: Proceedings of the Genetic and Evolutionary Computation Conference (2011)
27. Liapis, A., Yannakakis, G.N., Togelius, J.: Adapting models of visual aesthetics for personalized content creation. *IEEE Transactions on Computational Intelligence and AI in Games* **4**(3), 213–228 (2012)
28. Liapis, A., Yannakakis, G.N., Togelius, J.: Constrained novelty search: A study on game content generation. *Evolutionary Computation* **23**(1), 101–129 (2015)
29. Machado, P., Romero, J., Nadal, M., Santos-del Riego, A., Correia, J.a., Carballal, A.: Computerized measures of visual complexity. *Acta Psychologica* **160**, 43–57 (2015). <https://doi.org/10.1016/j.actpsy.2015.06.005>
30. Marcel, S., Rodriguez, Y.: Torchvision the machine-vision package of Torch. In: Proceedings of the ACM International Conference on Multimedia (2010). <https://doi.org/10.1145/1873951.1874254>
31. Michalewicz, Z.: Do not kill unfeasible individuals. In: Proceedings of the 4th Intelligent Information Systems Workshop (1995)
32. Mouret, J.B., Clune, J.: Illuminating search spaces by mapping elites. arXiv preprint arXiv:1504.04909 (2015). <https://doi.org/10.48550/arXiv.1504.04909>
33. OpenAI: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023). <https://doi.org/10.48550/arXiv.2303.08774>

34. Pugh, J.K., Soros, L.B., Stanley, K.O.: Quality diversity: A new frontier for evolutionary computation. *Frontiers in Robotics and AI* **3**, 40 (2016)
35. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *International Conference on Machine Learning*. pp. 8748–8763. PMLR (2021)
36. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. <https://openai.com/research/better-language-models> (2019), accessed 11 Jan 2024
37. Reimers, N., Gurevych, I.: Sentence-BERT: sentence embeddings using Siamese BERT-networks. In: *Proceedings of the Empirical Methods in Natural Language Processing Conference* (2019)
38. Ritchie, G.: Some empirical criteria for attributing creativity to a computer program. *Minds and Machines* **17**, 76–99 (2007)
39. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022)
40. Roziere, B., Teytaud, F., Hosu, V., Lin, H., Rapin, J., Zameshina, M., Teytaud, O.: EvolGAN: Evolutionary generative adversarial networks. In: *Proceedings of the Asian Conference on Computer Vision* (2021)
41. Secretan, J., Beato, N., D’Ambrosio, D.B., Rodriguez, A., Campbell, A., Stanley, K.O.: Picbreeder: evolving pictures collaboratively online. In: *Proceeding of the SIGCHI conference on Human factors in Computing Systems* (2008)
42. Sfikas, K., Liapis, A., Yannakakis, G.N.: Monte Carlo Elites: Quality-diversity selection as a multi-armed bandit problem. In: *Proceedings of the Genetic and Evolutionary Computation Conference* (2021)
43. Sfikas, K., Liapis, A., Yannakakis, G.N.: A general-purpose expressive algorithm for room-based environments. In: *Proceedings of the FDG workshop on Procedural Content Generation* (2022)
44. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: *Proceedings of the 32nd International Conference on Machine Learning* (2015)
45. Takagi, H.: Interactive evolutionary computation: Fusion of the capabilities of EC optimization and human evaluation. *Proceedings of the Institute of Electrical and Electronics Engineers* **89**(9), 1275–1296 (2001)
46. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: LLaMA: open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023). <https://doi.org/10.48550/arXiv.2302.13971>
47. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C.C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P.S., Lachaux, M.A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X.E., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., Scialom, T.: LLaMA 2: open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023). <https://doi.org/10.48550/arXiv.2307.09288>

48. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Proceedings of the Neural Information Processing Systems Conference (2017)
49. Viana, B.M.F., Pereira, L.T., Toledo, C.F.M.: Illuminating the space of enemies through MAP-Elites. In: Proceedings of the IEEE Conference on Games (2022). <https://doi.org/10.1109/CoG51982.2022.9893621>
50. West, P., Lu, X., Holtzman, A., Bhagavatula, C., Hwang, J.D., Choi, Y.: Reflective decoding: Beyond unidirectional generation with off-the-shelf language models. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (2021). <https://doi.org/10.18653/v1/2021.acl-long.114>
51. Xie, S., Tu, Z.: Holistically-nested edge detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2015). <https://doi.org/10.1109/ICCV.2015.164>
52. Zammit, M., Liapis, A., Yannakakis, G.N.: Seeding diversity into AI Art. In: Proceedings of the International Conference on Computational Creativity (2022)
53. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018). <https://doi.org/10.1109/CVPR.2018.00068>