Measuring Diversity of Game Scenarios

Yuchen Li, Ziqi Wang, Qingquan Zhang, Jialin Liu

Abstract—This survey comprehensively reviews the multi-dimensionality of game scenario diversity, spotlighting the innovative use of procedural content generation and other fields as cornerstones for enriching player experiences through diverse game scenarios. By traversing a wide array of disciplines, from affective modeling and multi-agent systems to psychological studies, our research underscores the importance of diverse game scenarios in gameplay and education. Through a taxonomy of diversity metrics and evaluation methods, we aim to bridge the current gaps in literature and practice, offering insights into effective strategies for measuring and integrating diversity in game scenarios. Our analysis highlights the necessity for a unified taxonomy to aid developers and researchers in crafting more engaging and varied game worlds. This survey not only charts a path for future research in diverse game scenarios but also serves as a handbook for industry practitioners seeking to leverage diversity as a key component of game design and development.

Index Terms—Game scenario, game level, game map, diversity, procedural content generation.

I. INTRODUCTION

VIDEO games have become a ubiquitous and influential media of entertainment, captivating millions of players and developers worldwide [1], [2]. Within the gaming world, one essential aspect that impacts player engagement and satisfaction is the diversity of video game content [3], [4]. As the domain of games continues to expand, there's a concerted move among researchers and designers towards devising effective methods to evaluate and amplify the richness of game content. This quest for diversity spans several domains, such as affective modeling, procedural content generation (PCG), game-based learning, play therapy, and trustworthy artificial intelligence (AI), each recognizing the unique contributions of varied game environments to their respective fields.

For instance, in the domain of affective modeling [3]–[7], the pursuit of diverse gaming content is driven by the goal to capture a broader range of player emotions, thereby maximizing player satisfaction through tailored emotional engagements. Similarly, in the field of multi-agent systems, there's a strong emphasis on enriching the diversity of scenarios, recognizing that the datasets for researching and validating trustworthiness AI is deeply connected to the diverse scenarios [8]–[11]. Psychological research delves into diverse game content to craft gameplay experiences that are not only more immersive but also educational, aiming to enhance the pedagogical efficacy of educational games [12]–[15].

Among this multidisciplinary pursuit, PCG emerges as a pivotal innovation [3], [4], [7], [16]–[21], revolutionizing game development by automating the creation of engaging content and thereby enhancing the player experience. The evolution of PCG reflects a paradigm shift from focusing solely on content quality to embracing a broader dimension of content diversity [4], [22], [23]. Initially concentrated on generating high-quality content, such as playable levels and challenges, the field of game AI is now exploring new horizons where diversity is as crucial as quality itself [2]–[4], [17], [22]. This transition acknowledges the increasing player demand for novel and varied experiences within games, urging developers to weave diversity into the fabric of game content [3], [22].

This transformation extends beyond the confines of academic research, resonating deeply with industry practices. Historical forays into PCG by pioneering games [3], [4], [24] like *Akalabeth* and *Rogue* lay the foundational stone for the Rogue-like genre, influencing countless successors and mainstream titles such as *Diablo*, which introduces players to randomized dungeons and item drops. The subsequent advent of commercial design tools marks a new era in PCG, shifting the burden of content creation from manual labor to automated processes [3], [4], [7], [16]–[21], as exemplified by modern classics like *Elite Dangerous* and *Minecraft* [3], [4]. These industry milestones highlight the versatility and profound impact of PCG across various gaming genres, underscoring its critical role in delivering diverse content [24], [25].

As we dive into the current situation of diversity measures, the lack of a systematic metric review of the metrics and evaluation methods used in the gaming industry and gaming research becomes evident. Researchers and developers often design their own diversity measures for their specific needs. There's an absence of a standardized approach. A further challenge is the absence of an overall taxonomy of the current diversity metrics and methods. Developers and researchers need to understand the category and characteristics of each metric to choose which metrics are most effective in different contexts, needs, and for various games.

Focusing on game scenarios, our survey aims to provide a systematic review and a taxonomy of diversity measures, helping researchers better understand diversity metrics and methods for their specific game scenarios and needs. By analyzing the existing diversity metrics and their applications, our survey seeks to create a roadmap for future works. We also provide valuable insights into these metrics discussing the multi-dimensionality of diversity measures. Additionally, we aim to guide researchers and developers in selecting the most appropriate representations, diversity metrics and methods for their specific game scenarios and game aspects. Ultimately, our survey aims to bridge the gaps in the current understanding of diversity metrics and provide a robust foundation for future research such as metric validation and comparison, and game development endeavors.

The structure of the paper is as follows. Section II outlines the methodology employed in this survey. Section III presents a taxonomy. The representation of game scenarios is examined in Section IV. This leads into Section V, which focuses on objective evaluation of diversity. Specifically, Section V-A summarizes the metrics used for the objective evaluation on these representations, and Section V-B delves into various objective evaluation approaches. In contrast, Section VI shifts the focus to subjective evaluation methods. The multi-dimensionality of diversity is discussed in Section VII. Section VIII discusses the implications and findings of these evaluations, and gives an outlook. Finally, Section IX concludes the paper.

II. METHODOLOGY OF SURVEY

Several key search terms have been used to guide our exploration of metrics and approaches for measuring diversity of video game scenarios. Inspired by [3] and [4], these terms include words such as "diversity", "novelty", "similarity", "expressive", "divergence", "diverse", and "duplication", which have been combined with terms related to "game" and "game metric". These selected keywords serve as the compass in navigating the vast sea of academic and industry literature. To ensure the comprehensiveness and reliability of our research, the literature review is conducted using a range of reputable databases and sources, including Scopus, ProQuest, IEEE Xplore, ACM Digital Library, and Google Scholar.

Though there exists a variety of content types [3], our survey focuses on game scenarios. In this research, the term "*scenario*" is specifically defined to include all interactive and non-interactive elements that contribute to the gameplay experience in video games. This encompasses the game's scenario, map, level, mechanics, and narrative elements that players engage with directly or indirectly during play. We distinctly exclude non-digital aspects, such as traditional board games and audio-visual elements like sound and music. Moreover, our scope does not cover in-game entity characteristics, such as characters, races, behaviors, costumes, the game's overall difficulty level, terrain unless evaluated within a game context, and the scope of affective computing concerning game diversity. For comprehensive coverage of terrain generation, readers are directed to detailed surveys by Voulgaris *et al.* [26], Galin *et al.* [27], Raffe *et al.* [28], and Valencia-Rosado *et al.* [29]. Similarly, while Kelly and McCabe's exploration of city generation features and techniques is acknowledged [30], it is considered minimally relevant to the concept of game scenarios as defined in our survey and is thus not included. Finally, discussions on dynamic difficulty adjustment are available in a survey by Mortazavi *et al.* [31].

Complementing these resources, Kutzia and von Mammen survey procedurally generated buildings [32]. Viana and Santos systematically reviewed dungeon generation methods [33]. However, to the best of our knowledge, none of these works mentioned evaluation metrics. Similarly, Gravina *et al.* propose PCG through quality-diversity but rarely discussed the specific evaluation methods for it [22]. A. Liapis [34] review a decade's research trends in PCG, noting a clear increase in the publication of evaluations in PCG papers. Yet, the majority of these papers still lack thorough evaluation [34]. Yannakakis and Melhart's systematic review on affective computing in games intersects with our interest but from a distinct angle focused on affective computing and player modeling [7], which diverges from our core focus on diversity metrics in game scenarios. In [25], a list of papers that used the expressive range to evaluate PCG generators is summarized. However, these evaluation metrics only apply to expressive ranges. While various metrics have been proposed and used, a structured taxonomy for diversity evaluation remains a need, as echoed in the works by [4], [16], [35].

In this survey, our focus is drawn towards the evaluation of video game items and the multi-dimensional nature of diversity in game scenarios through the lens of diversity metrics. While the broader discussions of social diversity within video games, touching upon aspects such as cultural nuances and representation, have been extensively explored in existing literature [36]–[39]. For those interested in the diversity of character identity within video games, the work by To *et al.* [40] is recommended. Our investigation delves into the diversity of game scenarios and components. This includes the variety and complexity of game elements, mechanics, and narrative elements that collectively shape the gameplay experience. By concentrating on the diversity of in-game elements and their assessment through various metrics [22], [25], we aim to contribute to the understanding of how diverse game scenarios can be quantitatively evaluated and categorized. This exploration does not directly address how diversity enhances player experiences but rather establishes a foundation for future research to link these quantitative assessments with qualitative player experiences, thereby enriching the dialogue around the role of diversity in video games.

III. TAXONOMY

In the context of PCG and game design, this survey categorizes diversity evaluation into two primary categories – *objective evaluation* and *subjective evaluation* – offers a structured approach to assessing the diversity of game scenarios. This bifurcation not only simplifies the understanding of how diversity can be measured but also highlights the different lenses through which the effectiveness of game scenarios can be evaluated. Figure 1 illustrates our taxonomy.

Objective evaluation in game scenarios is characterized by its reliance on quantifiable and directly measurable metrics without simulating gameplay. This category is further divided into *single-* and *multi-indicator* approaches, each offering unique insights into the diversity of content.

Single-indicator approaches focus on using a singular metric to evaluate diversity, within which, methods can be further classified into *comparison-based* and *non-comparison-based methods*. The metrics utilized by comparison-based methods are referred to as *comparison-based metrics*, and focus on evaluating the diversity of game scenarios by comparing elements against one another. This comparison can illuminate the variance within a set of generated items, highlighting how each piece of content differs from the others. In contrast, metrics of non-comparison-based methods, *non-comparison-based metrics* in short, analyze content independently of

Diversity Measure

Objective Evaluation (Sec. V)

Subjective Evaluation (Sec. VI)

Single-indicator (Sec. V-B1)

Multi-indicator (Sec. V-B2)

Human evaluation (Sec. VI-A)

Comparison-based (Sec. V-B1a)
○ Average Distance/Divergence
○ Average Nearest Neighbor Distance
○ Distribution of Similarity Values
○ Dissimilarity Map
○ Relative Diversity
○ Novelty Score
○ Computational Surprise
○ Slacking A-clipped Function
○ t-SNE Plot

Non-comparison-based (Sec. V-B1b)
○ Standard Deviation
○ Coefficient of Variance
○ Behavior Diversity
○ Spatial Diversity
○ Simpson Index

Expressive range (Sec. V-B2a)

Quality diversity (Sec. V-B2b)

Multi-objective (Sec. V-B2c)

Biological data (Sec. VI-B)
○ Facial/Head Expression
○ Heart Rate
○ Blood Volume Pulse
○ Skin Conductance
○ Borg Rating of Perceived Exertion

○ User Experience Questionnaire
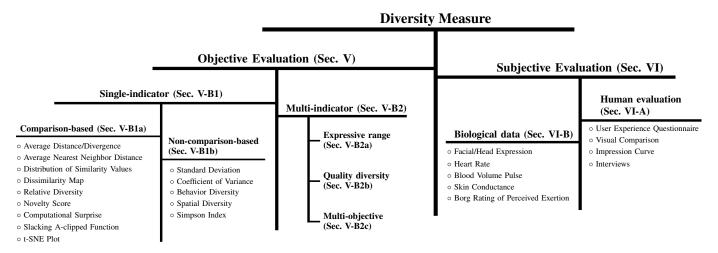○ Visual Comparison
○ Impression Curve
○ Interviews

Fig. 1. Taxonomy of measuring diversity of game scenarios.

others, assessing diversity based on the presence, absence, or magnitude of specific features within each item. Single-indicator approaches, while straightforward, can sometimes offer a limited perspective on diversity due to their focus on a singular aspect of content.

To capture a more holistic view of diversity, *multi-indicator approaches* combine several metrics, including those used in single-indicator approaches. This category of approaches is enriched by the inclusion of *visualization plots* and *cumulative scores*, uniquely suited to the multi-indicator analysis including expressive range analysis and quality-diversity analysis. Visualization plots can illustrate the distribution and relationships of various features within the scenarios such as quality-diversity maps [41] and expressive ranges [42], while cumulative scores aggregate multiple metrics into a single, comprehensive one such as quality-diversity score [41] and coverage [43]. This layered approach allows for a more nuanced understanding of scenario diversity, offering insights that might be missed when relying on a single indicator alone.

Subjective evaluation, in contrast to objective evaluation, involves human evaluation and biological data to assess the diversity of game scenarios. This method typically utilizes questionnaires, or biological data to gauge human responses to game content. Subjective evaluations capture the nuanced and sometimes intangible aspects of diversity that objective metrics might overlook, such as emotional impact, aesthetic appeal, and personal preference. However, subjective evaluation can be influenced by individual biases and is often more resource-intensive due to the need for participant involvement.

Both objective and subjective evaluations play crucial roles in understanding the diversity of game scenarios. The former offers quantifiable and reproducible metrics, while the latter provides insights into human perception and experience. Together, these approaches offer a comprehensive review for assessing and enhancing the diversity of game scenarios, ensuring that game elements not only vary significantly from one another but also resonate with users on a personal level.

Before delving into the specifics of various diversity evaluation approaches and metrics, Section IV reviews the representations of game scenarios upon which diversity is measured.

IV. REPRESENTATION OF GAME SCENARIOS

Game scenarios are originally represented as pixel-based images to users. Nonetheless, based on our comprehensive review, diversity assessments rarely leverage pixel-based representations directly [44], [45]. The representation of game components within game scenarios serves as a crucial foundation for measuring the diversity. Our survey captures the essence of the most widely-used representations in diversity measure and delves into two aspects: the definition of game components and their subsequent representation in game scenarios through *content-centered* and *player-centered* perspectives.

A. Definition of game components

We focus specifically on both empty and interactive components within games – elements that offer substantial content, map, functionality, or purpose beyond mere audio elements that don't directly contribute to gameplay or narrative. These critical components include objects, levels, and game mechanics that players can interact with or that significantly influence the game's progress or experience. Notably, while actions such as character decisions, item selections, strategy formulations, and decision-making processes indeed contribute to the richness of a player's experience, their diversification is categorized under a different facet of diversity, referred to as *policy diversity* in our research. This aspect, which delves into the strategic and decision-making diversity within the gameplay, is further discussed in Section VIII-C, distinguishing it from the direct measures of game components.

B. Content-centered vs. player-centered representations

With a clear understanding of the game components as defined above, this paragraph outlines two main types of commonly used representations to evaluate diversity in game scenarios: content-centered representation and player-centered representation. These representations are crucial for understanding and analyzing the dynamics and structure of games from both the content that comprises the game scenario and the interactions or behaviors of the players within that scenario. Figure 2 illustrates different representations.
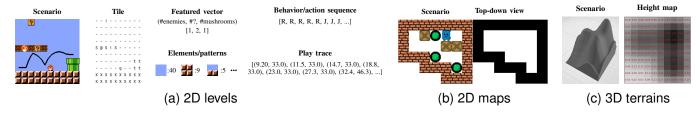
Fig. 2. Illustrative examples of game scenario representations.

Content-centered representation focuses on the structural aspects of the game. It includes various elements. *Elements/patterns* [43], [46] refer to the basic building blocks and the arrangement of those blocks within the game space, respectively. These can define the aesthetic and functional aspects of the game world. Strings, which might represent textual or data-based elements within the game's code or narrative structure. *Top-down view* [47] is a kind of representation used in the 2D or 3D game scenarios, highlighting the geometry/shape of game space, especially. Some games apply the top-down view first and then utilize tiles to represent their levels/rooms [15]. *Tiles* [43], [48]–[51] indicate the spatial design elements of games, such as the construction of levels or maps through discrete units or "tiles". Sequential cells [49], [50], suggesting a focus on the progression or order of tiles within a game, are often related to a more narrow-level design or structure in tiles. *Height maps* can represent terrains [28], [52], [53]. By recording the heights of a terrain at different coordinates, a complex terrain can be expressed by a 2D array. *Featured vectors* [17] represent some specific game elements in a low-dimensional mathematical space.

Player-centered representation, on the other hand, concentrates on the player's interaction with the game. *Behavior/action sequences* [54] document the series of actions or decisions made by players during gameplay, providing insights into player strategies, preferences, and challenges encountered. Notably, these sequences are closely related to edit distance calculations, which measure the dissimilarity between them. *Play traces* [55], [56] are records of player activities within the game, capturing the detailed paths players take, the choices they make, and the outcomes of those choices.

Both types of representations are essential for a comprehensive understanding of games, offering insights from the perspective of content creation and design, as well as from the player's interaction with and within the game environment.

V. OBJECTIVE EVALUATION

Objective evaluation includes a methodical analysis aimed at quantifying the diversity in video game scenarios. This critical examination is structured into two pivotal subsections: metrics for objective evaluation (cf. Section V-A) and its approaches (cf. Section V-B)[1]. The former introduces and categorizes a range of metrics designed to assess game content objectively. Following this, the objective evaluation approaches subsection

---

[1]Due to page limit, the metrics and methods for objective evaluation are formulated using unified notations in Supplementary Material.

transitions from the evaluation of isolated components to the methods of measuring diversity.

A. Metrics for objective evaluation

Within the domain of metrics for objective evaluation, the categorization of comparison-based metrics and non-comparison-based metrics offers a systematic approach to quantify scenario diversity. Comparison-based metrics provide a measure of variation between game components, highlighting the uniqueness of content, while non-comparison-based metrics delve into specific attributes, characterizing the detailed features that contribute to a game's diversity. This subsection details the metrics for the objective evaluation of diversity. It is worth mentioning that multiple metrics can be used simultaneously, which adopts a comprehensive perspective on diversity by amalgamating various metrics. This methodology is enhanced through the integration of visualization plots and cumulative scores, elements particularly conducive to analyzing diversity through multiple indicators, and detailed methods will be described in Section V-B2.

1) Comparison-based metrics: Comparison-based metrics are computed on representations of at least two game scenarios or through self-comparison, either vector-based or sequence-based representations (cf. Section IV). Included in the dissimilarity assessment are distance measures, divergence measures, and measures based on self-comparis employed to summarize scenario diversity.

a) Distance measures: A number of distance-based metrics have been used to compare a pair of game scenarios as they are simple yet straightforward.

One straightforward distance measure is **cosine similarity/distance** [57]. This metric is particularly effective when dealing with vector-based representations.

In the context of measuring the distances between representations, **Manhattan distance**, also known as the **L1 distance**, is particularly applicable when comparing vector-based or sequence-based representations of different scenarios. Oberberger *et al.* apply this metric to measure the distance between potential fields of tactics for wargaming and real-time strategy games [58]. They further use this metric in an evolutionary algorithm to search for a diverse set of tactics.

Euclidean distance, also known as **L2**, is also widely applied in comparison-based diversity methods [53], [59]–[62]. Preuss *et al.* define the **objective-based diversity** and **visual-impression diversity** as average nearest neighbor distance with different representations of levels [17]. The objective-based diversity evaluates several fitness functions to form a

vector representation of game maps, while the visual impression diversity extracts visual featured vectors. Both applied Euclidean distance as the comparison-based metrics.

Hamming distance is a distance measure commonly employed when dealing with discrete sequences of equal length, such as binary strings [17], [59], [63]–[69]. It calculates the number of positions at which the corresponding items between two sequences differ. It can also deal with 2D or 3D levels or maps by flattening the object into 1D discrete sequences.

Compression distance is a metric to measure the dissimilarity of two strings [70]. It is based on a compression algorithm (e.g., gzip), and is applied to compare 2D levels [23], [71].

Edit distance, also known as **Levenshtein distance**, is a metric employed to quantify the dissimilarity between two strings or sequences [72]. It measures the minimum number of single-character edits – insertions, deletions, or substitutions – required to transform one string into the other. In the context of comparing strings or sequences, the edit distance offers insights into the minimal sequence of edits needed to align or transform one into the other. This metric is applied to measure the dissimilarity of action sequences [23], [71], event sequences [73] and flattened 3D maps [74].

Constrained continuous edit distance is a variation of the traditional edit distance that introduces constraints on the allowed edit operations [75]. In the context of constrained edit operations, each operation (insertion, deletion, substitution) is assigned specific constraints or costs, allowing for a more nuanced assessment of dissimilarity based on the application's requirements. Osborn and Mateas develop *Gamalyzer*, a game-independent metric based on some refinements to constrained continuous edit distance to compare play traces [56]. The subsequent research by Osborn *et al.* confirm that *Gamalyzer* aligns more closely with human perception of "dissimilarity" and "uniqueness", concepts akin to diversity, and claimed that *Gamalyzer* can help investigate "*Do players pursue diverse strategies?*" [55].

Dynamic time warping (DTW) measures the similarity between two sequences that may vary in wrapping policy. Unlike traditional distance measures, DTW allows for the elastic alignment of time series, accommodating temporal distortions. DTW is applied to compare play traces [76].

N-Gram similarity and distance introduce a concept of measuring similarity and distance based on n-grams [54]. This approach demonstrates that traditional metrics like edit distance and the length of the longest common subsequence are special cases of n-gram distance and similarity, respectively. N-grams, which are contiguous sequences of n items (e.g., characters or words), offer a versatile framework for evaluating the similarity or dissimilarity between sequences.

b) Divergence measures: Sometimes, the distance between two game scenarios can be biased and is not able to correctly present the diversity of one or more game scenarios. Instead, comparing to a set of game scenarios is needed.

Similarity is a measure of how similar a content is compared to content in another content set [77]–[79]. Typically, the feature used in this metric is the count of elements.

Kullback-Leibler (KL) divergence, also called **relative entropy**, is the expectation of the logarithmic differences between two probability distributions. Lucas *et al.* define the tile-pattern distribution and apply KL divergence to compare the tile-pattern distributions or paired levels [80]. Tile-pattern distribution calculates the frequency of each distinct tile pattern in a level/map, allowing comparison between levels with the KL divergence metric. This metric is later adopted in comparing Super Mario Bros. (SMB) levels [81] and Minecraft maps [74].

Jensen–Shannon (JS) divergence, can be viewed as a symmetrized and smoothed variant of the KL divergence. Wang *et al.* apply JS divergence to compare tile-pattern frequency distributions of SMB levels [76].

Jaccard distance is commonly employed to compare dissimilarity between two sets of samples and it ranges from 0 to 1, with 0 indicating complete similarity (no difference between sets) and 1 indicating complete dissimilarity (no common items between sets). In the context of comparing sets of representations, Jaccard distance provides a robust measure of distance based on the relative sizes of the intersection and union of the sets. This metric is applied to measure the number of different winning strategies in the work by Nam *et al.* [62].

c) Measures based on self-comparison: A special case of comparison-based metric is comparing sub-sections of one single game scenario.

Symmetry, as the name suggests, is a measure of how symmetrical a level/map is along both the horizontal and vertical axes. Horizontally, it is computed by looking at pairs of rows starting at the center and moving outward and summing up the number of row positions that have the same tiles/content. Similarly, this is computed using pairs of columns for vertical symmetry [77]–[79]. The final symmetry value for a level/map is the sum of the horizontal and vertical symmetry.

d) Summary: Distance-based metrics such as Euclidean, Manhattan, and Hamming distances along with more complex measures such as cosine similarity and KL divergence, JS divergence, provide multiple choices for analyzing content diversity. The existence of this range of metrics showcases their ability to capture different aspects of diversity, from simple positional differences to distributional diversity. However, the abundance of distance-based metrics also reflects underlying challenges. First, no such single metric universally captures all dimensions of diversity, so a multi-dimensional approach is necessary. Secondly, the choice of metrics is influenced by the data structure, and also significantly influences the perceived diversity, as each metric emphasizes different aspects of the game scenarios. For instance, while Euclidean distance might highlight spatial diversity, KL divergence could focus on the differences in the distribution of game elements. These different metrics, therefore underscores the complexity of measuring game scenario diversity and point to an ultimate challenge: selecting the appropriate metrics that align with specific diversity dimensions of interest.

For games with spatially rich scenarios, such as platformer games, metrics that capture element diversity are essential. For instance, Manhattan distances can be applied effectively to measure the distance between game elements or levels. These distance-based metrics are particularly useful in vector representations. On the other hand, for behavior-specific

diversity, edit distance or DTW provides a more detailed assessment of diversity by capturing the changes along with behavior sequences or play traces. Understanding the nuances that different distributions of game elements bring to player experiences requires specific metrics, such as KL or JS divergence. These tools help highlight how varied the gameplay can be, based on how game elements or levels are distributed. Additionally, for games aiming to offer a visually diverse experience, using symmetry as a metric can effectively measure how aesthetically varied a game scenario is.

In conclusion, while different comparison-based metrics enrich our objective evaluation of game scenario diversity, it also highlights the trade-offs between comprehensiveness, specificity, and practicality. The challenge appears not just in the selection of metrics but also in capturing the multi-dimensional nature of diversity within game scenarios.

2) Non-comparison-based metrics: In non-comparison-based methods, the emphasis shifts from measuring the dissimilarity between items to extracting meaningful features or characteristics from individual objects. Non-comparison-based metrics are crucial in capturing relevant information for assessing diversity without relying on direct comparisons. The non-comparison-based metrics employed in assessing game-scenario diversity are summarized as follows.

In the analysis of the game scenario, the **counting number** metric signifies the total quantity of distinct components, objects, or entities within a specific context. Such a simple metric captures a wide array of items that significantly contribute to the composition and intricacy of the game scenario, such as the total number of jumps within a level and the subset of jumps deemed meaningful. A jump qualifies as meaningful if it is necessitated by the presence of an enemy or a gap, thus adding to the strategic complexity of the game [49]–[51]. The utility of the counting number metric extends across various domains, encompassing the evaluation of winning strategies [62], tracking alterations in knowledge graphs [82], quantifying unique player actions [83], identifying distinctive features within datasets [84], and the enumeration of unique local patterns [85]. Moreover, it is instrumental in analyzing the extremities of feature values, such as their maximums or minimums [46], thereby offering a comprehensive view of the game's diversity and complexity. Furthermore, several studies have employed the duplication percentage or the extent to which levels are replicated as a means to assess diversity across game levels, generators [66], [86]–[88], and within a broader dimension of games [89]. Regarding terrain generation in games, Frade *et al.* group similar terrains and count the number of groups as a diversity measure [90]. They further introduce two metrics: the accessibility score based on the count of the inaccessible area and the obstacles edge length score [90], [91]. Subsequently, they combine these metrics and develop a method that incorporates the weighted sum of these two metrics to evaluate terrain programs based on their height maps [52], [92]. The counting number metric is further utilized as a key evaluative evaluation method in numerous studies. In [93], Szilas and Ilea delve into the assessment of diversity in interactive narratives, distinguishing between intra-diversity and global diversity. Intra-diversity refers to the count of differ-ent actions within a single playthrough, emphasizing session-specific diversity. Conversely, global diversity measures the array of unique actions across multiple playthroughs, thereby gauging the overall diversity encountered by players [93]. This metric has also been utilized in [94]–[96].

Linearity is a measure of how well the level geometry aligns with a straight line, determined through linear regression. Each level is scored by summing the absolute values of the distances from each point per unit to its expected value on the line, normalized by either the maximum linearity value or the maximum number of units [42], [48]. In some cases, a unit may refer to the center point of each platform [42], or every column in a level [48].

Nonlinearity is a measure based on the linear regression error when fitting a line to the structures within a segment, employing a similar formulation to linearity [43]. Notably, this metric is also applied to dungeon levels, where mission linearity captures the linearity of the mission structure and map linearity assesses the linearity of map layouts [97].

Leniency is often calculated based on the quantity of specific items within a unit, such as enemies, gaps, or safety features, contributing to the game's difficulty or ease [42], [98]. Variations in calculating leniency might include the total number of enemies and gaps, offset by the number of rewards, providing a nuanced view of the level's challenge [49]–[51].

Density is a metric that quantifies the frequency of specific items within a level, such as the number of tiles that are neither background nor path tiles [43], solid blocks [48], and others. Density metrics in diversity evaluations can include the percentage of completable levels [49]–[51], empty space [49]–[51], reachable space [49]–[51], interesting tiles among others [49]–[51], and useless rooms [97].

a) Summary: Non-comparison-based metrics offer a unique way to examine game scenarios by focusing on their individual features, such as the number of enemy types in a level or the complexity of story choices. They provide a detailed picture of what each game scenario offers without direct comparison. This approach raises an interesting point: even though these metrics start by examining scenarios individually, the insights they provide can lead to comparisons and their specific characteristics. For example, observing how strategies vary across levels or how narrative choices diversify gameplay. Essentially, those metrics also highlight characteristics of an individual level. When applied to a broader collection of levels, this metrics collection can better highlight a specific set designed to cater to player preferences.

B. Objective evaluation approaches

In the context of objective evaluation approaches, the distinction between single-indicator approaches and multi-indicator approaches marks a pivotal methodology in assessing game scenario diversity. Single-indicator approaches focus on employing a singular metric to evaluate a specific aspect of diversity, offering a targeted and straightforward assessment method. In contrast, multi-indicator approaches adopt a more holistic view by integrating various metrics to provide a comprehensive analysis of diversity across multiple dimensions.

1) Single-indicator approaches: Within single-indicator approaches, we distinguish between comparison-based methods and non-comparison-based methods, each offering unique perspectives on evaluating game scenario diversity. Comparison-based methods concentrate on quantifying the dissimilarities between game elements or scenarios, leveraging mathematical metrics to measure the distance or difference between them. On the other hand, non-comparison-based methods focus on the intrinsic characteristics of individual game components or scenarios without directly comparing them to others.

a) Comparison-based methods: One fundamental approach for assessing diversity is through comparison-based methods, which calculate the distance or dissimilarity between contents and employ statistical techniques to access diversity measures. The outcome of these methods typically expresses a scalar value signifying the extent of diversity.

Average distance/divergence is a natural and widely-used metric to measure diversity [23], [73], [99]. A notable portion of literature concerning tile-based levels or maps utilizes average Hamming distance to evaluate the scenario diversity. For example, Earle *et al.* propose controllable PCG via reinforcement learning (RL) and computes average Hamming distance [99]. Jiang *et al.* extend this method to the 3D map generation of Minecraft, while the average Hamming distance is inherited as the diversity measure [69]. In works for online PCG on the SMB benchmark [59], [64], [67], average Hamming distance is applied as the scenario diversity measure. Zook *et al.* evolve tactical field care scenarios and evaluate the scenario diversity of their designed generator by computing edit distance between all scenarios in the evolution population, and illustrate the average and max distance over the by curves [73]. Mariño *et al.* suggest average compression distance is a proper measure of structural diversity [100]. Later, Beukman *et al.* also utilize average compression distance, along with average edit distance on agent's action sequence, to evaluate the diversity of SMB levels [23], [71]. Awiszus *et al.* also apply average edit distance to evaluate diversity, however, they directly compute the distance on the flattened string of 3D Minecraft maps [74], supplemented by an average tile-pattern KL divergence. Nam *et al.* represent role-playing game stages via event-parameter vectors, and employed average distance to evaluate the diversity of stages generated by RL policies [62]. Two distance metrics are used in this work, to instantiate two ad-hoc diversity measures. The first is Euclidean distance which directly applies to the vector representation of stages. The second is the number of different winning strategies in two stages, which can be viewed as a variant of the Jaccard distance without normalization. To investigate the limitation of RL-based generators in online level generation, Wang *et al.* evaluate the average Euclidean distance of *latent vectors* [59]. Wulff-Jensen *et al.* calculate mean square error and structured similarity index between randomly sampled pairs of generated 3D game maps, and report the average, median, standard deviation, and standard error of dissimilarity values [53].

Preuss *et al.* propose the **average nearest neighbor distance** to measure the diversity of sets of 2D game maps [17]. This method calculates the average of the distance between set items and its nearest neighbor within a reference set. This method is adopted by Wang *et al.* to evaluate the diversity of level slices represented by *latent vector* [59]. Instead of summarizing the similarity value into a scalar value, Khalifa *et al.* show the **distribution of similarity values** via a curve plot of the probability density [101]. This method could reveal more underlying information regarding diversity.

Another method with similar motivation, but different visualization is to plot the dissimilarity values with a heatmap, namely a **dissimilarity map**. Marczak *et al.* calculate a matrix of similarity values in gameplay audio of two players [102]. This matrix is visualized through a heatmap, allowing for the visualization of both similar and dissimilar experiences between the players during gameplay. Through this approach, the matrix effectively highlights the degree of interaction and experience overlap or divergence between the participants. Schubert *et al.* calculate the distances between generated levels and human-crafted levels and visualize the distances in a 2D heatmap [81]. These works compare the dissimilarity between two sets of scenarios, rather than the dissimilarity within a set.

Coman and Muñoz-Avila define **relative diversity** by addressing the average distance between a scenario and each scenario in a set, to conduct case-based reasoning for playing real-time strategy games [103], [104].

Novelty score is usually used as a fitness function in evolutionary algorithms [60]. It measures the contribution to the population diversity of an individual. Similarly, Gravina *et al.* define **computational surprise** as the behavioral difference between an item and its corresponding prediction [105].

Lehman and Stanley introduce *novelty search* as an approach to discovering behavioral novelty [60], [61], initially tested in maze environments. Subsequently, Liapis *et al.* apply novelty search with constraints in the field of PCG [63], [106]. The evaluation in novelty search relies on the novelty score, which calculates the average distance between an individual and its closest neighbor, as detailed in [63], [106]. In their later research [68], they refer to Hamming distance as **visual diversity**, which involves comparing two maps on a tile-by-tile basis, presenting it as the comparison-based metric in novelty score. Later, Gravina *et al.* introduce *surprise search*, which utilizes a **surprise metric** based on Euclidean distance to measure the behavioral dissimilarity between an individual and its expected behavior [105]. Subsequently, they also introduce a weighted sum of the novelty score and surprise score, referred to as the **local competition score** compared with novelty score and surprise score [107]. Sudhakaran *et al.* propose MarioGPT to generate game levels from text prompts with a combination of large language model and novelty search, which employ the novelty score to evaluate the diversity [108]. Novelty score is also applied with metrics such as tile-pattern KL divergence [109], [110], Hamming distance [111], edit distance [112], [113] and compression distance [71].

Slacking A-clipped function is a method evaluating the diversity of content within a scenario [76]. It is devised for assessing online generated level segments, rewarding the moderate divergences of a new level segment regarding several previously generated ones. Wang *et al.* embed both the tile-pattern JS divergence as a content comparison metric and the

DTW as a gameplay comparison metric into this function to evaluate multi-facet diversity in SMB levels [76].

The **t-distributed stochastic neighbor embedding** (t-SNE) is a famous method to embed high-dimensional data into low-dimensional space, with the distance relationship approximately preserved [114]. t-SNE is mostly used to embed the data into 2-dimensional space low-dimensional and visualize the embedded points via scatter plots. Though this plot is not directly related to diversity, by plotting different sets of game scenarios in the same figure, their diversities can be compared.

Snodgrass and Ontañón train level generators with only 25 training levels [115]. The generated levels are embedded with training levels and visualized. The plot shows that generated levels form a wide-spread distribution covering several training levels, indicating the generated levels are not overfitted but maintained diversity. Similarly, Schubert *et al.* generate levels with generative adversarial networks and visualize them along with training levels through the t-SNE plot [81]. Wang *et al.* plot different sets of level slices with the t-SNE plot, and advocate the one with the most complex patterns may have the best diversity [59]. Moreover, t-SNE is used to compare the "exploration" of novelty search and random sampling [108].

b) Non-comparison-based method: Non-comparison-based methods often rely on metrics or counts related to specific characteristics or specific items within a game scenario. This could involve assessing the distribution of a particular item or attribute within the game scenario. The output of non-comparison-based methods is also usually a scalar value in single-indicator approaches, but these non-comparison-based metrics can also be combined in multi-indicator approaches (cf. Section V-B2) with other formats of outputs.

If the scenarios are represented as a single feature value, it is applicable to assess diversity via the **standard deviation** [116]–[120]. A vanilla standard deviation method is widely applied in multiple game scenarios with selected non-comparison-based metrics. Cook *et al.* introduce analytical techniques for evaluating generator samples, emphasizing two essential aspects: the centroid, representing the average *connectedness and density* score, and the standard deviation of the *connectedness score* with a same initial solid chance parameter setting, which indicates the dispersion of the sample [121]. Wang *et al.* report the standard deviation bullet hell game barrages' ("Danmaku") *shooting frequency*, *mean momentum*, and *(screen) coverage* [116]. They claim that possibly a larger standard deviation indicates better diversity.

Furthermore, Zhang *et al.* design a **coefficient of variance** measure to evaluate the *gameplay event diversity* while playing a game level, which is defined as the fraction of standard deviation and the mean value [122]. Sorochan *et al.* utilize four metrics for evaluating the different aspects of generated levels, which include *gold total per level*, *percentage collected per level*, *total nodes explored*, and *nodes per gold*. They also compute the standard deviation of each of these metrics across the generated levels. To visualize the results, they employ a boxplot for standard deviation and histograms to represent the distribution of interesting tiles and space within the game Lode Runner [123].

Another widely used method, **behavior diversity**, is originally proposed to measure "diversity in opponents behavior over the games" based on the standard deviation by Yannakakis *et al.* [119], [124]. Later this method is also adopted in other games with distinct non-comparison-based metrics [118]. In addition to behavioral diversity, **spatial diversity** is another crucial aspect explored by Yannakakis *et al.* [119], [124]. This method focuses on the configuration and composition of the game scenario. By segmenting a map into various feature values, such as nodes, tiles, cells, and other spatial elements, researchers can quantify the spatial diversity of a game scenario. The calculation of spatial diversity employs *Shannon entropy*, a measure derived from information theory that provides a statistical means of evaluating the unpredictability or complexity of a system [118], [124]–[127]. Typically, the counting number is employed as the non-comparison-based metric in spatial diversity.

In their works of 2005 and 2007 [119], [120], [124], Yannakakis and Hallam introduce various quantifications of interest criteria, with a specific focus on two key diversity aspects: behavior diversity and spatial diversity. Behavior diversity is quantified by calculating the standard deviation of the *time taken*, while spatial diversity is measured using the entropy of *visited cells*. These metrics are applicable to a broad range of predator/prey computer games, as they are based on generic features common to this category of games. Later, the concept of spatial diversity measured by entropy of *visited nodes* is also adopted in research involving "fun" analysis by Yannakakis *et al.* [127]–[129]. Pedersen *et al.* compute the spatial diversity based on the gap count and visualize the count of each collected feature of player behaviors to evaluate the diversity of generated levels [130]. Later, Sombat *et al.* also use spatial diversity of *maze features* and behavior diversity of the *survival duration* to model interest in gameplay [118].

In the 2010 Mario AI Championship [131], the concept of **spatial diversity** is adopted widely during this event. They evaluate diversity on multiple levels based on *gaps* and *enemy placements*. Furthermore, they collect statistics on the *numbers of coins, rocks, and powerups* as essential features to compare competitors in the championship, highlighting the importance of various gameplay elements in assessing diversity. Togelius *et al.* advocate that spatial diversity reflects the interestingness of StarCraft maps [125]. Loiacono *et al.* apply extracted frequency distribution from curvature profiles and speed profiles, to establish two diversity measures for the track of high-end racing games *entropy* [132]. Dutra *et al.* define the diversity of a scenario as the entropy of event occurrences [133]. They use entropy as both the reward function to train RL-based level generators and the evaluation criterion.

Simpson index [134] is a classic measure of diversity by addressing the number of categories and element frequencies which is similar to spatial diversity. It ranged in $[0, 1)$. The more uniform the frequency distribution is, the higher the Simpson index value. This measure is adopted by Berland *et al.* to evaluate both game choice diversity and game outcome diversity [135].

2) Multi-indicator approaches: Another main branch of diversity evaluation methods is the multi-indicator approach. A scenario is projected into a multi-dimensional feature space,

where each dimension corresponds to a single diversity indicator. The diversity is reflected by the distribution of the scenario in that feature space. We delve domain of multi-indicator approaches into expressive range analysis (ERA), quality-diversity (QD) analysis, and multi-objective analysis, each presenting a sophisticated lens through which game scenario diversity can be examined. The output of these methods can be scalar values indicating the degree of diversity or charts visualizing the diversity.

a) Expressive range analysis: ERA is a widely adopted method for comprehending complex generative spaces by transforming intricate high-dimensional representations of potential artifacts into more accessible 2D visualizations [42], [136], [137]. ERA visualizes sets of levels in different scopes by employing two or more measurable metrics mentioned in Section V-A. Utilizing these metrics, each artifact is localized within a 2D plot as expressive ranges, providing a concise overview of how the generated content is distributed in this metric-defined space [137]. The visualization can be implemented by a **scatter plot** showing each artifact's location in the metric-defined space, or by a **height map** with its grid corresponding to a partition of the feature space. The color of each cell in the height map represents the number or frequency of the scenarios in its corresponding subspace. By condensing the richness of generative spaces into visually interpretable representations, ERA facilitates a more intuitive understanding of the characteristics and variations.

Building on Smith *et al.*'s groundwork for evaluating game level [136], their later work introduces an **expressive range scatter plot**, focusing on *linearity* and *leniency* values for each level, thereby refining the analysis within specified ranges and deepening the understanding of the diversity in level set [98], [138]. This also marks the emergence of a standardized method for ERA, which employs weighted hexagons to visualize how levels distribute across different parameters, offering insights into the diversity of game [42]. Further advancements are made with the introduction of geometric similarity, utilizing edit distance on vectors encoding geometry types for each beat, providing a nuanced scale for assessing diversity [139], [140].

This method underscores the ongoing interest in measuring the expressivity of a generator especially and, by extension, the diversity of generated levels, even though the term "*diversity*" is not explicitly used in their conventional sense. The emphasis on "*expressivity*" as a means to evaluate generator diversity is further validated in the book [3], where the authors clarify that visualizing expressivity acts as a method to assess the diversity of the generator. This focus on expressivity and the subsequent development of evaluation methods illustrate a nuanced approach towards understanding and assessing the diversity of game levels.

Subsequent works have further embraced expressive range analysis for evaluating diversity, focusing on the statistically meaningful number of one or multiple considered objects, such as patterns/game elements [43], [49]–[51], [66], [77]–[79], [141], [142], color islands [143], and the density of considered objects [43], [48]–[51], [66], [69], [77]–[79], [141], [144]. Other key metrics are also applied in ERA, including linearity [25], [48]–[51], [77]–[79], [115], [141], [144]–[147], leniency [25], [48]–[51], [77]–[79], [115], [141], [144]–[147], graph complexity [148], [149], danger [148], [149], axiality [149], generated paths [66], [69], [142], [149], and the comparison of shortest to winning paths [66], [69], [149], along with symmetry [43], [77]–[79], [141] and similarity [43], [77]–[79]. In addition, several metrics including entropy [150], compression distance [141], and edit distance [139], [140], augment diversity assessments by providing comprehensive diversity scores.

The use of various visualization plots, including histograms [86], [123], [144], scatter plots [43], [97], [151]–[153], heatmap [22], [25], [49]–[51], [77]–[79], [110], [141], density estimation [51], [153], corner plots [44], [49]–[51], density plots [47], dissimilarity maps [141], raincloud plots [150], box plots [141], and codependency analysis [121], further enriches this analytical approach. A corner plot is typically a visualization used in statistics and data analysis to display the relationship between two metrics in a grid of scatter plots [51], [154]. For a more straightforward comparison, Cook *et al.* use expressive range to visualize changes in expressivity through the automatic optimization of generators, employing the tool *Danesh* for this purpose [155]–[157]. To select suitable metric pairs, Withington and Tokarchuk evaluate various metric pairs, ranking them to highlight the bias from improper metric selection [137].

Additionally, multiple video game genres are evaluated by ERA to measure diversity considering some genres-specific metrics. For room-based games, metrics like diameter, solidity, rooms, and skewness [44], [45], as well as the feasibility score and a multi-indicator approach including plan compactness and average room compactness for top-down room generation, have been utilized [47]. In Minecraft, 12 metrics, including light, defense, functional metric, aesthetic metric, etc., are considered, with some showing a notably high correlation with human evaluation scores [150]. Moreover, the challenges of 3D level generation have been tackled through specific tasks like maximizing level diameter, ensuring door connections, and generating dungeon levels [69].

b) Quality-diversity analysis: In the multi-indicator approach, QD algorithms represent a significant component, constituting a distinct category of evolution-inspired techniques. Their goal is to simultaneously maintain both the quality and diversity of solutions [158]. This is achieved by characterizing generated content, referred to as *behavior characterization*, via multiple metrics mentioned in Section V-A. The concept of behavior characterization plays a crucial role in driving the diversity component in QD analysis. These metrics could represent different features that evaluate various aspects of content [41], [159], especially for diversity considerations. The idea is to capture a diverse range of behaviors [107].

The **QD-score** serves as a conventional evaluation metric for QD algorithms. It is based on a partition of the feature space, aggregating the quality scores of the best content within each partitioned subspace [41].

The **coverage** measure calculates the proportion of subspaces occupied. This metric signifies the extent to which a specific algorithm such as MAP-Elites [43], successfully located solutions across the search space during the run.

A **QD map**, often referred to as a QD-space or behavior space, is a graphical representation that visualizes the quality evaluation of a diverse range of solutions through a heatmap. The colors in the heatmap correspond to the quality evaluation values [41].

In the early stages of QD research, evaluations typically rely on fitness values with *tile entropy* and *derivative-tile entropy* of the level, supplemented by the *expressive ranges* [22], [160]. This period marks the beginning of a broader exploration of game content and gameplay behaviors, encompassing diverse features within these domains.

For instance, the complexity of block and structural generation in Minecraft-inspired platforms is examined [161], while bullet hell games utilize diversity metrics such as entropy–calculated from the first, second, and third derivatives of an agent's action sequence–alongside risk, and distribution [162]. Subsequent studies extend this exploration to include game mechanics in games on general video game AI platform [163] and non-comparison-based metrics in platformer, puzzle, and dungeon games, with a focus on density/frequency [43], [48], [164], nonlinearity/linearity [43], [48], [77], [78], leniency [48], [164], symmetry [43], [65], [99], similarity [77], [78], presence of game elements [43], [65], [77], [78], [99], [164]–[166], solution/path length [65], [99], [166], presence of game actions [165], and the dynamics of game actions, such as speed, time, size, and jump entropy [167]. Moreover, KL divergence from ground-truth levels has been employed for evaluations [165].

In "Baba is You", a dynamic, rule-altering sokoban-style game, the rule activation distribution is depicted using distribution histograms [168]. In deckbuilding, Fontaine *et al.* introduce the QD algorithm, utilizing average mana and mana variation to visualize the distributions of deck performance and density distributions of deck populations [169]. Later, for the diverse strategies, the average hand size and the average number of turns, the max health differences are featured and visualized by a QD map [170], [171].

Typically, visualizations and cumulative assessments of diversity in QD research might encompass elements such as QD maps, which can feature heat maps [65], [165]–[167], [169]–[172], bar or histogram charts [162], [163], [168] and heat maps with density distributions [169]. The QD score [43], [65], [164], [165], [170]–[172], coverage metrics [43], [47], [77]–[79], [165], [172], and the percentage or number of filled cells/elites [65], [163], [164], [168], [170], [171] also play a significant role. Most of these works also visualize the evolution process of QD-score or coverage via curve plots [43], [164], [165], [170]–[172]. Specifically, Biemer *et al.* visualize the frequency of runs found bin for each generator through heatmaps, akin to the expressive range [48].

In 2019, Gravina *et al.* propose PCG through QD (PCG-QD) as a subset of search-based PCG [22]. In their work, each component of PCG-QD is well-organized with discussion, and for more information beyond game generation about PCG-QD, one can refer to this work.

c) Multi-objective analysis: Multi-objective optimization is a technique to find a wide range of solutions approximating the Pareto front, which is the optimal set of solutions that are not dominated by any possible solution. The "dominate" means a solution is not worse than the other in terms of any considered objective, while is better than the other in terms of at least one objective. There have been some works investigating multi-objective searches for game scenario generation, and the diversity in terms of objective values is widely concerned.

Those works typically plot the locations of generated scenarios in the objective space by a scatter plot, to illustrate the diversity, we refer to this as multi-objective scatter plot. In the context of StarCraft map generation, Togelius constructs five computational metrics for the consideration of fairness, interestingness, aesthetics, and playability, then applies multi-objective optimization to generate maps with those characteristics. Another work by Togelius *et al.* emphasizes the importance of skill differentiation and interestingness [173], both of which are closely tied to map diversity. They employ two fitness functions, choke points and path overlapping, to assess these aspects. In a later work [174], they also employ a fitness function related to resource clustering, which is similarly modeled using Shannon's entropy. Khalifa and Togelius conduct a comprehensive evaluation of their multi-objective level generator across various aspects [175]. In the case of Binary, a maze game, the evaluation involves calculating the number of regions and the improvement in path length within the generated map. All of those works illustrate the diversity through the multi-objective scatter plot.

Wang *et al.* train multiple generators with different trade-offs between fun score and diversity by varying a weight parameter [67]. They also illustrate the diversity of those scenario generators by the multi-objective scatter plot showing their locations in the objective space.

Hypervolume is a frequently-used performance measurement in multi-objective optimization, which evaluates the overall performance of a solution set in terms of convergence and diversity among the considered objectives. Hypervolume computes the volume enclosed by the solution set in the objective space. Besides the multi-objective scatter plot, the aforementioned work by Togelius *et al.* also calculate hypervolume. Ma *et al.* incorporate the modeling approach in the work by Togelius *et al.* [174] and propose a new multi-objective optimization algorithm to generate MegaGlest maps [126], supplemented by hypervolume.

VI. SUBJECTIVE EVALUATION

Subjective evaluation stands as a pivotal method for understanding the diversity within game scenarios, providing an in-depth look at how players perceive and interact with various elements of a game. This feedback is crucial for developers, illuminating both the strengths of the game and areas ripe for improvement. It ensures that the game not only offers a wide range of experiences but also resonates with a diverse audience. It fosters a deeper understanding of scenario diversity from the player's perspective, guiding future enhancements to better satisfy varied player needs. Typically, subjective evaluation measures a player's diverse experiences within a single game as an assessment of diversity in scenarios, but it can also extend to a horizontal comparison across different games within the same genre.

A. *Human evaluation*

Human evaluation plays a crucial role in assessing the diversity of game scenarios through qualitative assessments and annotations conducted by human evaluators. This method, inherently subjective, offers deep insights into the nuanced, subjective aspects of diversity that automated metrics might overlook. Common approaches to human evaluation include various subjective assessments, and biological data annotations, providing a comprehensive understanding of the qualitative dimensions of game diversity.

Using a **user experience questionnaire (UEQ)** to measure game scenario diversity during or after play provides critical insights into player perceptions and engagement. Questions targeting scenario variety, challenges, and interaction dynamics help assess the game's narrative and gameplay diversity. This method collects subjective feedback, offering a detailed view of the game's appeal across different player preferences. The questionnaire can be enriched with human annotations like *rank-based evaluation*, *pairwise comparison*, and other evaluation methods, further deepening the analysis. Osborn *et al.* investigate the relation between their proposed "dissimilarity metric" and humans' perception of dissimilarity, outliers, and uniqueness by a UEQ on a 7-Likert scale [55]. Similarly, Saurik *et al.* implement a UEQ on a 7-Likert scale to estimate the novelty of a horror game [176]. UEQ is also used in [177] with a similar process on a mobile-based game. Hämäläinen *et al.* propose a framework that encompasses five facets for evaluating movement-based games *after the interaction* [178]. The paper is structured around five central, gravity-related facets of user experience, identified based on their work and that of others: realism, affect, challenge, movement diversity, and sociality. Special attention is given to the aspect of movement diversity. Partlan *et al.* propose an evaluation method for interactive narratives *during the interaction*, utilizing four key graphical representations: the scene graph, layout graph, script graph, and interaction maps [179]. This evaluation method encompasses a playthrough of a scenario, an examination through a visual script editor, and a detailed walk-through of these representations. Each walk-through concludes with a set of questions aimed at scrutinizing the narrative structure, the scenario's interaction with the player, and other pertinent aspects. Notably, the second walk-through specifically focuses on visualizations provided by the script graphs and interaction maps, offering a deeper insight into the representation. The testers will naturally notice the aspects of diversity in game scenarios, and the designers can also lead them by asking diversity-related questions. However, Szilas and Ilea [93] suggest that an in-game questionnaire helps to better understand the player experience but may interrupt the overall experience.

Visual comparison serves as an intuitive and immediate method for evaluating game scenario diversity, offering a direct way to observe differences by presenting a set of samples. This approach allows stakeholders, including developers, players, and researchers, to visually assess the variety and richness of game elements side by side. By examining these samples, one can easily identify the range of scenarios, interactive narratives, and gameplay mechanics, highlighting the diversity within a game or across different games. Visual comparison not only makes the assessment of diversity straightforward but also provides concrete examples that can support discussions and decisions regarding game design and improvement. This technique is particularly effective in conveying the nuances of diversity that might be overlooked in textual or numerical evaluations, bringing a clear, impactful perspective on the visual and experiential variety present in game scenarios. In the work of Sarkar *et al.* [43], examples of levels under different settings are presented to show diversity visually. Especially, they encode the absence or presence of the corresponding element as an N-digit binary number with the total N of element types considered for the specific game. The generated levels with N-digits and other characteristics are displayed in [43] for readers to *visually compare* the diversity of each level. Similarly, Steckel and Schrum also demonstrate the *visual diversity* of their generated levels in their work [166]. Sudhakaran *et al.* plot the play traces of an agent on generated levels to illustrate the gameplay diversity [108]. Holmgård *et al.* present images of a level with different markers indicating different players' researched positions, to showcase the diversity of player decision-making styles [180].

The **impression curve**, introduced by K. Wejchert [181], [182], is a concept that explores how space, time, and motion influence an observer's perception, particularly in the context of spatial images within different interior layouts, such as a street sequence. This method illustrates how various elements within a space impact an observer over time, without a specific measure but rather through a subjective scale from 1 to 10, where 1 represents a lack of architectural value and 10 signifies strong, meaningful architectural features. According to Andrzejczak *et al.* [182], the impression curve has been applied in various studies to analyze and evaluate the diversity and significance of different spaces, including urban streets and rural landscapes, by mapping the observer's changing impressions over time.

Conducting **interviews** provides a unique opportunity to gain in-depth insights into player experiences, perceptions, and preferences regarding game scenario diversity. This qualitative method facilitates a direct conversation between researchers and players, allowing for a comprehensive exploration of the nuanced ways in which different game elements are received and interpreted by individuals. Interviews can uncover detailed feedback on specific scenarios, reveal player strategies, and highlight emotional responses that other evaluative methods might not capture. Schrum *et al.* conduct a user study to investigate the experience of using their proposed interactive game design tool [183]. Despite the diversity is not a mandatory question, they receive some comments regarding diversity.

B. *Biological data*

Measuring the diversity of a video game using biological data presents a novel approach to understanding the complexity and variety within game scenarios, and narratives. By applying methodologies for analyzing biological data, such as heart-rate changes, one can quantify the range of elements

and interactions in a game. This could involve examining the variety of available strategies, or the complexity of the game's world. Such an analysis can provide insights into the richness of the gaming experience, offering a unique perspective on how well the game simulates varied real-world scenarios or accommodates different player strategies.

Facial or head expressions are critical non-verbal communication cues that convey a wide range of emotions, thoughts, and intentions without the use of words. Facial expressions result from one or more motions or positions of the muscles beneath the skin of the face. Shaker *et al.* employ a range of features to assess player experiences, including *game level (content) features*, *gameplay behavioral features*, and *head movement features* [184]. Analyzing players' *head expressivity* in response to specific in-game events provides insights into the diversity of player gameplay experiences. To evaluate these experiences, they administer four alternative forced-choice questionnaires to players after they have played games with different feature values. These questionnaires allow players to rate their preferences for engagement, challenge, and frustration states. The use of *pairwise preferences* in this assessment minimized the influence of factors such as personality and culture on self-reported player experiences, as these emotions are often closely linked to diverse gameplay experiences. Elor *et al.* visualize players' performance, physiological responses, neural responses, and facial movements during both the foundation and challenge protocols [185]. These visualizations are complemented by an emotion survey, highlighting that the game sessions could evoke a diverse range of emotions.

Yannakakis *et al.* have conducted a series of studies aimed at modeling entertainment [5], [12], [127]–[129], [186], by analyzing children's physiological responses during play in physical interactive playgrounds inspired by computer games. These studies focus on capturing and modeling children's affective states, particularly through measurements of **heart rate** (HR), **blood volume pulse** (BVP), and **skin conductance** (SC), during physical gameplay. Yannakakis *et al.*'s research, which utilizes these physiological signals to model entertainment values, provides a foundational basis for exploring the relationship between game diversity, physiological responses and "fun" experiences [5], [12], [127]–[129], [186]. By selecting game benchmark levels based on diversity measures, one aligns with Malone's principles of intrinsic qualitative factors—challenge, curiosity, and fantasy–that are essential for engaging gameplay. The variability, statistics and dynamic changes in HR, BVP, and SC during diverse gameplay scenarios can offer insights into how different game elements trigger varying levels of engagement and emotional arousal. For instance, games that introduce unexpected challenges or novel scenarios might elicit stronger physiological responses, indicating heightened curiosity or excitement, while those offering rich, fantasy-driven contexts could engage players on deeper emotional levels. Therefore, by analyzing the correlations between physiological responses and the diversity of game scenarios, researchers can better understand how variations in gameplay contribute to the overall entertainment experience, offering a quantitative measure to supplement traditional qualitative assessments of game diversity.

HR and the **borg rating of perceived exertion (RPE) scale** are pivotal metrics for evaluating the intensity and physical exertion of players in virtual reality (VR) games that involve physical exercise during their gameplay. Yoo *et al.* conduct a study where they track players' percentage of maximum HR alongside the RPE for multiple VR games [187], presenting their findings in a plot that also incorporates questionnaire responses.

VII. Multi-dimensionality of Diversity Evaluation

In the dynamic world of games, the pursuit of diversity in game scenarios presents a multi-dimensional challenge that touches upon a comprehensive and detailed approach, transiting from simple numerical evaluations to more complex and subjective approaches.

A. Which representation to use?

The selection of appropriate representations for measuring diversity is directly influenced by the facets under consideration. There exists a clear mapping between facets and their corresponding representations, which plays a crucial role in the accurate assessment of game scenario diversity. For facets related to player interaction, as referred to player-centered representation in Section.IV, such as play traces [55], [56], [76], action/behavior sequences [23], [71], the chosen representations are inherently tied to the temporal and behavioral aspects of gameplay. Conversely, content-related facets are represented through content-centered representations such as tiles [43], [48]–[51], top-down views [47], featured vectors [17], and content sequences [123], reflecting the static or structural elements of the game environment.

The choice of metrics for evaluating diversity is significantly informed by these representations. For instance, DTW is a comparison-based metric closely associated with temporal sequences, making it particularly suited for analyzing player behavior, where the timing and order of actions carry meaningful information [76]. On the other hand, metrics for content-related diversity, such as tile-based KL divergence [80], and Hamming distance [17], [63], focus on comparing static features or configurations within game levels. These content-based metrics excel in quantifying the variation in level design, content features, or the presence of specific game elements.

This approach to selecting representations and corresponding metrics underscores the multi-dimensional nature of diversity in game scenarios. By tailoring the analytical framework to the specific facets of diversity, researchers can derive more meaningful and actionable insights into how diversity manifests within both the player experience and the game content. As the field progresses, the exploration of new representations and metrics suitable for untapped facets of diversity, such as music or narrative structures, promises to further enrich our understanding of what makes game environments engaging and diverse. This evolution in diversity measurement methodology highlights the importance of a comprehensive and flexible approach, capable of adapting to the diverse and dynamic nature of games.

B. Which evaluation approach to choose?

In objective evaluation approaches, especially, diverse methodologies are employed and diversity can be categorized into three dimensions: *intra-diversity*, *inter-diversity*, and *overall diversity*. This structured framework reflects the various perceptions of diversity among researchers who held different research interests, aiming to explore the multi-dimensional nature of game scenarios.

1) Intra-diversity: It focuses on the diversity present within individual components, levels, maps, or sessions of a game [73], [76]. This category of measures illuminates the range of game elements, or options available in a singular context, scenario, level, or map, akin to assessing the variety of actions within a single period of gameplay or the unique elements within a specific level. The goal is to capture the richness and depth offered in isolated segments of gameplay, providing insight into the moment-to-moment variety that players might encounter.

2) Inter-diversity: Inter-diversity, on the other hand, examines the diversity across different components or sessions [17]. This perspective considers the variance between multiple playthroughs, levels, or game modes, offering a broader view of the diversity present in the game as a whole. It's about understanding how each part contributes to a diverse gaming experience, ensuring that players encounter a wide range of scenarios as they progress through the game.

3) Overall diversity: Overall diversity is employed as a holistic value or an overall visualization that captures the entirety of a game's diversity. This metric and/or diagram aim to quantify the game's total diversity, merging both the intra- and inter-diversity perspectives into a comprehensive evaluation. It serves as a summary indicator of the game's ability to offer players varied and engaging experiences, reflecting the combined effect of all diversity aspects.

By incorporating these diverse metrics–intra-diversity, inter-diversity, and overall diversity–researchers can provide a multi-dimensional analysis of game scenarios. This methodology acknowledges the layered and subjective nature of diversity, aligning more closely with the intricate ways in which players experience and appreciate variety in gaming. The distinction between the scopes of diversity, coupled with an overall evaluation, highlights the complexity of game design and the importance of a systematic approach to measuring diversity. For instance, Zook *et al.* utilize the edit distance to evaluate a single scenario and then compute the average and maximum value representing both intra- and inter-diversity for their generated scenarios, supplemented by a plot of diversity value over iterations for the overall diversity [73].

C. Which methodology to choose?

This section discusses in a logical sequence that mirrors an analytical progression of diversity evaluation, illustrating appropriate metric pair dimensions toward different interests.

a) Multi-facet diversity metrics: At the outset, the examination of multi-facet diversity introduces multiple metrics involving different facets of the game considering both player behavior and content to evaluate game scenarios. Yannakakis

has led the way in assessing diversity through measures like behavioral diversity and spatial diversity [12], [120], [124], [128]. These metrics, focusing on the variations in opponent behaviors and the game content [12], [120], [124], [128], highlight the richness of diversity within individual game elements. This methodical approach provides insights into the layers of game content, emphasizing the significance of each game element's contribution to the overall gaming experience.

b) Multi-dimensional diversity metrics: Building on the foundation established by facet-specific metrics, the methodology extends to the multi-dimensional implementation of metrics. Unlike the previously discussed multi-faceted metrics, this approach pairs metrics implemented in varied relationships, such as within individual levels and between different levels. This multi-dimensional methodology includes a metric implemented in levels of a generator and in levels of multiple generators [23], [141], a metric implemented in generated levels and in-between generated levels and training/target levels [87], [88], [112], multiple metrics implemented in pairs of level and in addressing the expressivity of a generator [69], or multiple metrics considering diversity within generator and across multiple generators [67]. Through the applications of these multi-dimensional approaches, these works illuminate the different approaches to comprehensively understanding diversity within specific dimensions, highlighting the multi-dimensional nature of diversity in game scenarios.

c) Visualization of diversity metrics: Finally, the synthesis and portrayal of game scenario diversity through the visualization of metrics mark the culmination of this structured methodology. The use of expressive ranges and curves for comparative analysis, either against training data benchmarks or within generator outputs in the context of PCG, is instrumental. These visual representations are vital in articulating the extent and depth of diversity across various systems, seamlessly connecting theoretical concepts with their practical applications in game design.

Undoubtedly, measuring diversity in game scenarios is a multi-dimensional and multi-faceted endeavor. While it remains unclear which method can precisely capture the essence of game diversity, there is a hopeful anticipation for diverse perspectives to illuminate the concept of diversity from various angles. This aspiration towards a multi-dimensional understanding encourages continuous exploration and dialogue within the field, fostering a richer comprehension of how diversity shapes the gaming experience.

VIII. DISCUSSION AND OUTLOOK

This section addresses several pivotal considerations, starting with the quest for the optimal degree of diversity within game scenarios. We also tackle the critical need for robust validation methods for diversity evaluation metrics, advocating for empirical studies that correlate these metrics with actual player experiences and preferences. The interplay between policy diversity and scenario diversity is examined, highlighting how AI strategies and diverse game scenarios influence the overall gaming experience and AI trustworthiness. Moreover, the influence of game genre on diversity evaluation is discussed,

emphasizing the unique challenges and opportunities presented by different genres in achieving diverse game content. Finally, we identify and discuss gaps in current research and practice, including the disparity between academic theories and industry application, the potential of AI in automating diversity evaluation, and the need for interdisciplinary approaches that span content, strategy, player behavior, and emotional diversity. This section not only illuminates the challenges at hand but also reveals the opportunities for bridging gaps between theoretical frameworks and practical applications.

A. What is the desired diversity degree?

The examination of game scenario diversity in the literature often goes beyond mere quantification. Many studies prioritize the evaluation of dissimilarity between generated and predefined target game levels, sometimes addressing both dissimilarity and diversity. Todd *et al.* highlight the importance of balancing similarity and dissimilarity within generated and target sets, suggesting that a moderate level of dissimilarity can be beneficial [112]. Some other research focuses on closely aligning generated levels with human-created examples or training datasets, (e.g., [15], [188], [189]).

Zakaria *et al.* examine specific in-game objects, such as the levels' *entropy*, *empty tile percentage*, and *idle crates frequency*, to analyze how dissimilar procedurally generated levels are from those designed by humans [66]. Similarly, Torrado *et al.* explore differences between generated and human-designed games by visualizing the count of levels alongside average Hamming distance and the diversity of game tiles [86]. Wang *et al.* propose a metric addressing intra-diversity of each level at a moderate value, engaging a fun experience [76]. Siper *et al.* introduce another perspective on diversity [87], distinguishing between inter-diversity, which measures the dissimilarity between generated levels and goal-set levels, and intra-diversity, which considers the uniqueness within each generated map. This differentiation is instrumental in understanding the complexity of diversity in game design and reflects the range of approaches within the field.

These varied approaches each offer unique contributions towards the collective goal of creating game scenarios that are engaging, diverse, and well-balanced. Through the lens of these studies, the field navigates the complex dynamics between generated content and predefined benchmarks, exploring both dissimilarity and alignment with human-designed ones.

B. Validation of diversity evaluation

Exploring the correlation between diversity metrics and their relevance across various gaming contexts has become a crucial area of research, aiming to bridge the divide between academic findings and practical applications within the game development industry. This area of research not only assesses the impact of varied game scenarios on player enjoyment and satisfaction [5], [39], [120], [127], [190], [191], but also emphasizes the influence of these metrics on enhancing player experiences. Insights derived from such analyses are critical in creating immersive and engaging gaming environments that meet or exceed player expectations.

Further complicating this field is the observation that traditional metrics like the mean squared error and the structural similarity index can yield contrasting results when comparing content generated by different algorithms [53]. This discrepancy underscores the challenge of selecting appropriate metrics and suggests that a singular metric may be insufficient for a comprehensive analysis of game level diversity.

Despite the growing body of work exploring the relationship between diversity metrics and player perception, the direct analysis of this interplay remains relatively unexplored. Most existing user studies have concentrated on the capacity of these metrics to evoke player emotions, rather than on their ability to foster diverse content or experiences. Notably, the research by Osborn *et al.* confirms that *Gamalyzer* aligns more closely with human perception of "dissimilarity" and "uniqueness", concepts akin to diversity [55]. Fontaine *et al.* [165] conduct a study assessing the perceived similarity between generated levels and human-designed levels, aiming to determine if the KL divergence can model perceived similarity effectively. Other research extends similar investigation into the context of player emotions [100], [118], [120], [192], underlining the significance of aligning diversity metrics with player perceptions to create more engaging and satisfying gaming experiences. This evolving discourse highlights the need for a comprehensive approach that considers both the objective and subjective dimensions of game diversity, emphasizing the critical role of player feedback and advanced analytics in shaping the future of game design.

C. Policy diversity & scenario diversity

Many games–especially in *real-time strategy games*, *fighting games*, and *role-playing games*, involve the interaction between players and agent-controlled characters. Though the diversity of agent behaviors is beyond our scope of "scenario diversity", sometimes it can crucially affect the game experience. Therefore, we briefly discuss some measures regarding the diversity of agent behaviors as follows.

There is a wide range of multi-agent RL research proposing different diversity measures for agent behaviors. For example, Vinyals *et al.* plot the distribution of the average number of each unit built by Protoss agents throughout league training, normalized by the most common unit, to evaluate the agent [193]. Lupu *et al.* define the diversity of agent behaviors as the average JS divergence with respect to the decision-trajectory distribution, from the individual policies to the average policy [194]. Some research measures the diversity of agent behavior in a multi-objective sense. Zheng *et al.* visualize their trained game-playing policies through scatter plots with winning rate and exploration score as the two axes. The work by Shen *et al.* [195] presents multi-objective scatter plots, and further visualizes the gameplay trajectories of their agents trained for *Justice Online*.

Besides, comparison-based methods, QD analysis, and some other specialized methods are also employed to measure agent diversity. Szubert *et al.* [196] calculate the average Hamming distance between the action sequences of agents as a measure of behavioral diversity. Yuda *et al.* present

the minimum, maximum, and average cosine similarity of each agent's featured vector across all others to exhibit their behavior diversity [57]. Canaan *et al.* leverage QD search to generate diverse agents to play Hanabi [197]. Coverage and averaged QD-score are reported and the QD-maps are illustrated. They also proposed intra-run diversity and cross-run diversity to evaluate the performance of their algorithm in terms of finding behaviorally diverse agents. Halina and Guzdial propose to measure the diversity of an agent relative to others [198] as the percentage of random states that the agent takes a distinct action compared to each other. A higher distinct percentage identifies a more distinct game-playing policy. Gaina *et al.* tune game parameters via an evolutionary algorithm to minimize the performance of default agents and maximize the performance of specialized agents, assuming that higher fitness value indicates greater strategic diversity [199].

D. Game genre driven diversity evaluation

One notable aspect pertains to the evaluation of diversity within different game genres. A number of games or game-based platforms are available for researching AI and PCG [200]. While certain game genres have been extensively researched, others have received relatively limited attention. This discrepancy underscores the need to explore diversity evaluation methods tailored to specific genres, as what constitutes diversity may vary significantly between genres. For example, *first-person shoot (FPS)* games, known for their competitive nature, emphasize the balance of in-game resources and player-versus-player interactions. Consequently, the focus of PCG in these games is primarily on ensuring balanced gameplay, fairness in player confrontations, and difficulty rather than on the internal diversity of the maps [201], [202]. This approach contrasts with other game genres where map diversity might play a more central role. The lack of literature specifically addressing the internal diversity of maps in FPS games suggests that this aspect has not been a major focus in the field. This observation opens up opportunities for future research to explore how PCG can enhance the diversity of map design in FPS games, potentially enriching the player experience by offering varied environments and strategic challenges.

E. Gaps

The gaps existed in industry and academia, emerging technologies, cross-disciplinary, and educational contexts underscore substantial challenges in the cultivation and assessment of game diversity:

a) Industry vs. research in game diversity metrics: There's a notable disconnect between the theoretical frameworks developed in academia for measuring game diversity and their practical application within the industry. While academic circles propose sophisticated and novel methods for assessing diversity, the industry's uptake is often swayed by user feedback [176], [177] and gameplay data analytics by questionnaires [203]–[205], rather than rigorous academic models. This indicates a need for researchers to rethink the application of diversity metrics in game design, potentially by leveraging player data and feedback to guide the development of more relevant evaluation tools.

b) Large language models as evaluators: With the advancement of technology, AIs are playing an increasingly crucial role in automating the evaluation and enhancement of diversity within game scenarios. According to Gallotta *et al.*, RAGAS is one of the examples utilizing one large language model (LLM) to evaluate generated game content by other LLMs [206]. This presents an opportunity for the creation of real-time evaluation tools and self-regulating mechanisms for measuring diversity, based on player evaluations, to seamlessly integrate diverse content creation into the game development process.

c) Cross-disciplinary gaps: There is a vast gap between different fields of research in how they approach and measure diversity. For instance, research in multi-agent systems and RL might focus on the diversity of strategies and policy, while PCG emphasizes the diversity of game content. Affective modeling considers both content and player experience diversity but tends to categorize diverse psychological reactions into one or a few emotions. Moreover, very few studies utilize biological data to measure game diversity, despite its potential to provide a comprehensive understanding of diversity that includes content, strategy, player behavior, perceived diversity, and physical responses (e.g., exercise intensity or heart rate), especially relevant for VR or movement-based games [5], [120], [128], [129], [187]. Fields like serious game design and dynamic difficulty adjustment focus more on learning experience and game difficulty, often overlooking the other aspects of games [13], [31], [207].

d) Educational evaluation methods: Educational evaluation methods in the context of leveraging the whole student-designed games and simulations for learning often emphasize creativity and the development of computational thinking skills. Koh *et al.* have contributed significantly to this area by developing a methodology that bridges computational thinking with science simulations [14], [208], [209]. Their approach involves the use of cosine similarity and divergence scores (based on Euclidean distance) to compare student-created simulations against a standard sample. This innovative method not only assesses the creativity and technical accuracy of the simulations but also quantifies the extent to which students internalize and apply computational thinking principles. Additionally, techniques like the use of a similarity matrix to analyze students' creations reveal insightful patterns in the application of programming concepts [210]. This evaluation strategy underscores the critical relationship between creative endeavors and computational skills in educational settings, offering a nuanced perspective on student learning and achievement in digital simulation and game design.

Addressing these gaps requires a concerted effort to develop interdisciplinary approaches that encompass content diversity, strategy diversity, player behavioral diversity, and emotional diversity. This holistic view should ideally extend to measuring biological responses, thereby offering a more comprehensive assessment of game diversity suited for various types of games, including VR and movement-based games. By bridging these divides, researchers and developers can better understand and enhance the multi-dimensional nature of game diversity, ultimately leading to richer, more engaging gaming experiences

IX. CONCLUSION

In conclusion, our survey highlights the critical role of diversity in video game scenarios, directly impacting player engagement, satisfaction, and overall gaming experience. By examining a spectrum of evaluation methods and metrics, we illuminate the nuanced approaches to measuring game scenario diversity, revealing a complex interplay between content creation and player interaction. Our findings underscore the necessity of a unified taxonomy and a structured strategy for choosing diversity measures that can guide both academic research and practical application in game development.

The gaps identified between industry practices and academic research, alongside the emerging role of AI and LLMs in automating content generation and evaluation, present both challenges and opportunities. Bridging these gaps requires fostering closer collaboration between researchers and game developers, leveraging the strengths of both worlds to create more diverse, engaging, and innovative game experiences.

Looking forward, the continuous evolution of PCG technologies and methodologies offers a promising avenue for addressing the identified gaps and further advancing the diversity of game scenarios. Future research should focus on developing adaptive and dynamic evaluation metrics that can better capture the multi-dimensional nature of diversity in games. Moreover, exploring underrepresented game genres and incorporating player feedback and biological data into diversity evaluation could provide deeper insights into the subjective experience of diversity in gaming.

Embracing diversity in game scenarios not only enriches player experiences but also reflects the diverse interests and backgrounds of the global gaming community. As AI continues to play a pivotal role in shaping the future of game automation, concerted efforts toward understanding and enhancing game scenario diversity will be essential in creating more inclusive, engaging, and innovative gaming environments.

REFERENCES

[1] B. Jovanovic. (2023) Gamer demographics: Facts about the most popular hobby. https://dataprot.net/statistics/gamer-demographics/, last accessed on 07 April 2024.

[2] S. Risi and J. Togelius, "Increasing generality in machine learning through procedural content generation," *Nature Machine Intelligence*, vol. 2, no. 8, pp. 428–436, 2020.

[3] N. Shaker, J. Togelius, and M. J. Nelson, *Procedural Content Generation in Games*. Springer, 2016.

[4] G. N. Yannakakis and J. Togelius, *Artificial Intelligence and Games*. Springer, 2018.

[5] G. N. Yannakakis and J. Hallam, "Entertainment modeling through physiology in physical play," *International Journal of Human-Computer Studies*, vol. 66, no. 10, pp. 741–755, 2008.

[6] A. Liapis, G. N. Yannakakis, M. J. Nelson, M. Preuss, and R. Bidarra, "Orchestrating game generation," *IEEE Transactions on Games*, vol. 11, no. 1, pp. 48–68, 2019.

[7] G. N. Yannakakis and D. Melhart, "Affective game computing: A survey," *Proceedings of the IEEE*, vol. 111, no. 10, pp. 1423–1444, 2023.

[8] I. Sahin and T. Kumbasar, "Catch me if you can: A pursuit-evasion game with intelligent agents in the unity 3D game environment," in *2020 International Conference on Electrical Engineering*, 2020, pp. 1–6.

[9] Z. Liu, C. Yu, Y. Yang, P. Sun, Z. Wu, and Y. Li, "A unified diversity measure for multiagent reinforcement learning," in *Advances in Neural Information Processing Systems*, vol. 35. Curran Associates, Inc., 2022, pp. 10 339–10 352.

[10] C. Sun, S. Shen, S. Xu, and W. Zhang, "Diversity is strength: Mastering football full game with interactive reinforcement learning of multiple AIs," *arXiv preprint arXiv:2306.15903*, 2023.

[11] C. Guerrero-Romero, S. Lucas, and D. Perez-Liebana, "Beyond playing to win: Creating a team of agents with distinct behaviours for automated gameplay," *IEEE Transactions on Games*, vol. 15, no. 3, pp. 469–482, 2023.

[12] G. N. Yannakakis and J. Hallam, "Modeling and augmenting game entertainment through challenge and curiosity," *International Journal on Artificial Intelligence Tools*, vol. 16, no. 06, pp. 981–999, 2007.

[13] P. Wouters, E. D. van der Spek, and H. van Oostendorp, "Measuring learning in serious games: A case study with structural assessment," *Educational Technology Research and Development*, vol. 59, no. 6, pp. 741–763, 2011.

[14] K. H. Koh, V. Bennett, and A. Repenning, "Computing indicators of creativity," in *Proceedings of the 8th ACM Conference on Creativity and Cognition*. ACM, 2011, pp. 357–358.

[15] K. Park, B. W. Mott, W. Min, K. E. Boyer, E. N. Wiebe, and J. C. Lester, "Generating educational game levels with multistep deep convolutional generative adversarial networks," in *2019 IEEE Conference on Games*, 2019, pp. 1–8.

[16] G. N. Yannakakis and J. Togelius, "Experience-driven procedural content generation," *IEEE Transactions on Affective Computing*, vol. 2, no. 3, pp. 147–161, 2011.

[17] M. Preuss, A. Liapis, and J. Togelius, "Searching for good and diverse game levels," in *2014 IEEE Conference on Computational Intelligence and Games*. IEEE, 2014, pp. 1–8.

[18] A. Summerville, S. Snodgrass, M. Guzdial, C. Holmgård, A. K. Hoover, A. Isaksen, A. Nealen, and J. Togelius, "Procedural content generation via machine learning (PCGML)," *IEEE Transactions on Games*, vol. 10, no. 3, pp. 257–270, 2018.

[19] B. De Kegel and M. Haahr, "Procedural puzzle generation: A survey," *IEEE Transactions on Games*, vol. 12, no. 1, pp. 21–40, 2020.

[20] J. Liu, S. Snodgrass, A. Khalifa, S. Risi, G. N. Yannakakis, and J. Togelius, "Deep learning for procedural content generation," *Neural Computing and Applications*, vol. 33, no. 1, pp. 19–37, 2021.

[21] M. Guzdial, S. Snodgrass, and A. J. Summerville, *Procedural Content Generation via Machine Learning: An Overview*. Springer, 2022.

[22] D. Gravina, A. Khalifa, A. Liapis, J. Togelius, and G. N. Yannakakis, "Procedural content generation through quality diversity," in *2019 IEEE Conference on Games*, 2019, pp. 1–8.

[23] M. Beukman, C. W. Cleghorn, and S. James, "Procedural content generation using neuroevolution and novelty search for diverse video game levels," in *Proceedings of the Genetic and Evolutionary Computation Conference*. ACM, 2022, pp. 1028–1037.

[24] A. Amato, "Procedural content generation in the game industry," in *Game Dynamics: Best Practices in Procedural and Dynamic Game Content Generation*. Springer, 2017, pp. 15–25.

[25] R. Dolfe, "Mixed-initiative tile-based designer : Examining expressive range and controllability for 2D tile-based levels," Ph.D. dissertation, KTH Royal Institute of Technology, 2022.

[26] G. Voulgaris, I. Mademlis, and I. Pitas, "Procedural terrain generation using generative adversarial networks," in *29th European Signal Processing Conference*, 2021, pp. 686–690.

[27] E. Galin, E. Guérin, A. Peytavie, G. Cordonnier, M.-P. Cani, B. Benes, and J. Gain, "A review of digital terrain modeling," *Computer Graphics Forum*, vol. 38, no. 2, pp. 553–577, 2019.

[28] W. L. Raffe, F. Zambetta, and X. Li, "A survey of procedural terrain generation techniques using evolutionary algorithms," in *2012 IEEE Congress on Evolutionary Computation*. IEEE, 2012, pp. 1–8.

[29] L. O. Valencia-Rosado and O. Starostenko, "Methods for procedural terrain generation: A review," in *Pattern Recognition*. Springer, 2019, pp. 58–67.

[30] G. Kelly and H. McCabe, "A survey of procedural techniques for city generation," *The ITB Journal*, vol. 7, no. 2, pp. 87–130, 2006.

[31] F. Mortazavi, H. Moradi, and A.-H. Vahabie, "Dynamic difficulty adjustment approaches in video games: A systematic literature review," *Multimedia Tools and Applications*, 2024, doi: 10.1007/s11042-024-18768-x.

[32] D. Kutzias and S. von Mammen, "Recent advances in procedural generation of buildings: From diversity to integration," *IEEE Transactions on Games*, pp. 1–20, 2023.

[33] B. M. F. Viana and S. R. dos Santos, "Procedural dungeon generation: A survey," *Journal on Interactive Systems*, vol. 12, no. 1, pp. 83–101, 2021.

[34] A. Liapis, "10 years of the PCG workshop: Past and future trends," in *Proceedings of the 15th International Conference on the Foundations of Digital Games*. ACM, 2020, pp. 1–10.

[35] N. Shaker, G. Smith, and G. N. Yannakakis, "Evaluating content generators," in *Procedural Content Generation in Games*. Springer, 2016, pp. 215–224.

[36] Y. Sato, "Cross-cultural game studies," in *Encyclopedia of Computer Graphics and Games*. Springer, 2021, pp. 1–6.

[37] C. J. Passmore, R. Yates, M. V. Birk, and R. L. Mandryk, "Racial diversity in indie games: Patterns, challenges, and opportunities," in *Extended Abstracts Publication of the Annual Symposium on Computer-Human Interaction in Play*. ACM, 2017, pp. 137–151.

[38] A. Harvey, "Becoming gamesworkers: Diversity, higher education, and the future of the game industry," *Television & New Media*, vol. 20, no. 8, pp. 756–766, 2019.

[39] J. Weststar and M. Legault, "Developer satisfaction survey: Summary report [2021] – IGDA," International Game Developers Association, Tech. Rep., 2022.

[40] A. To, J. McDonald, J. Holmes, G. Kaufman, and J. Hammer, "Character diversity in digital and non-digital games," *Transactions of the Digital Games Research Association*, vol. 4, no. 1, pp. 32–65, 2018.

[41] J. K. Pugh, L. B. Soros, P. A. Szerlip, and K. O. Stanley, "Confronting the challenge of quality diversity," in *Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation*. ACM, 2015, pp. 967–974.

[42] G. Smith and J. Whitehead, "Analyzing the expressive range of a level generator," in *The 2010 Workshop on Procedural Content Generation in Games*. ACM, 2010, pp. 1–7.

[43] A. Sarkar and S. Cooper, "Generating and blending game levels via quality-diversity in the latent space of a variational autoencoder," in *Proceedings of the 16th International Conference on the Foundations of Digital Games*. ACM, 2021, pp. 1–11.

[44] E. Giacomello, P. L. Lanzi, and D. Loiacono, "Searching the latent space of a generative adversarial network to generate DOOM levels," in *2019 IEEE Conference on Games*, 2019, pp. 1–8.

[45] ——, "DOOM level generation using generative adversarial networks," in *2018 IEEE Games, Entertainment, Media Conference*, 2018, pp. 316–323.

[46] S. Beaupre, T. Wiles, S. Briggs, and G. Smith, "A design pattern approach for multi-game level generation," in *Proceedings of the Fourteenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*. AAAI, 2018, pp. 145–151.

[47] K. Sfikas, A. Liapis, and G. N. Yannakakis, "A general-purpose expressive algorithm for room-based environments," in *Proceedings of the 17th International Conference on the Foundations of Digital Games*. ACM, 2022, pp. 1–9.

[48] C. Biemer, A. Hervella, and S. Cooper, "Gram-elites: N-gram based quality-diversity search," in *The 16th International Conference on the Foundations of Digital Games*. ACM, 2021, pp. 1–6.

[49] A. Summerville and M. Mateas, "Super Mario as a string: Platformer level generation via LSTMs," *arXiv preprint arXiv:1603.00930*, 2016.

[50] A. Summerville, M. Guzdial, M. Mateas, and M. Riedl, "Learning player tailored content from observation: Platformer level generation from video traces using LSTMs," in *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 12, no. 2. AAAI, 2016, pp. 107–113.

[51] A. Summerville, "Expanding expressive range: Evaluation methodologies for procedural content generation," in *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 14, no. 1, 2018, pp. 116–122.

[52] M. Frade, F. F. de Vega, and C. Cotta, "Aesthetic terrain programs database for creativity assessment," in *2012 IEEE Conference on Computational Intelligence and Games*, 2012, pp. 350–354.

[53] A. Wulff-Jensen, N. N. Rant, T. N. Møller, and J. A. Billeskov, "Deep convolutional generative adversarial network for procedural 3D landscape generation based on DEM," in *Interactivity, Game Creation, Design, Learning, and Innovation*. Springer, 2018, pp. 85–94.

[54] G. Kondrak, "N-gram similarity and distance," in *String Processing and Information Retrieval*. Springer, 2005, pp. 115–126.

[55] J. Osborn, B. Samuel, J. McCoy, and M. Mateas, "Evaluating play trace (dis)similarity metrics," in *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 10, no. 1. AAAI, 2014, pp. 139–145.

[56] J. C. Osborn and M. Mateas, "A game-independent play trace dissimilarity metric," in *Proceedings of the 9th International Conference on the Foundations of Digital Games*. ACM, 2014, pp. 1–7.

[57] K. Yuda, S. Kamei, R. Tanji, R. Ito, I. Wakana, and M. Mozgovoy, "Identification of play styles in universal fighting engine," *arXiv preprint arXiv:2108.03599*, 2021.

[58] M. Oberberger, S. J. Louis, and M. Nicolescu, "Evolving team tactics using potential fields," in *2013 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, 2013, pp. 1–8.

[59] Z. Wang, T. Shu, and J. Liu, "State space closure: Revisiting endless online level generation via reinforcement learning," *IEEE Transactions on Games*, 2023, doi: 10.1109/TG.2023.3262297.

[60] J. Lehman and K. O. Stanley, "Abandoning objectives: Evolution through the search for novelty alone," *Evolutionary Computation*, vol. 19, no. 2, pp. 189–223, 2011.

[61] J. Lehman, K. O. Stanley, and R. Miikkulainen, "Effective diversity maintenance in deceptive domains," in *Proceedings of the 15th Annual Conference on Genetic and Evolutionary Computation*. ACM, 2013, pp. 215–222.

[62] S.-G. Nam, C.-H. Hsueh, and K. Ikeda, "Generation of game stages with quality and diversity by reinforcement learning in turn-based RPG," *IEEE Transactions on Games*, vol. 14, no. 3, pp. 488–501, 2022.

[63] A. Liapis, G. N. Yannakakis, and J. Togelius, "Sentient sketchbook: Computer-assisted game level authoring," in *Proceedings of the 8th International Conference on Foundations of Digital Games*. ACM, 2013, pp. 1–8.

[64] Z. Wang and J. Liu, "Online game level generation from music," in *2022 IEEE Conference on Games*. IEEE, 2022, pp. 119–126.

[65] S. Earle, J. Snider, M. C. Fontaine, S. Nikolaidis, and J. Togelius, "Illuminating diverse neural cellular automata for level generation," in *Proceedings of the Genetic and Evolutionary Computation Conference*. ACM, 2022, pp. 68–76.

[66] Y. Zakaria, M. Fayek, and M. Hadhoud, "Procedural level generation for Sokoban via deep learning: An experimental study," *IEEE Transactions on Games*, vol. 15, no. 1, pp. 108–120, 2023.

[67] Z. Wang, C. Hu, J. Liu, and X. Yao, "Negatively correlated ensemble reinforcement learning for online diverse game level generation," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=iAW2EQXfwb

[68] A. Liapis, G. N. Yannakakis, and J. Togelius, "Constrained novelty search: A study on game content generation," *Evolutionary Computation*, vol. 23, no. 1, pp. 101–129, 2015.

[69] Z. Jiang, S. Earle, M. Green, and J. Togelius, "Learning controllable 3D level generators," in *Proceedings of the 17th International Conference on the Foundations of Digital Games*. ACM, 2022, pp. 1–9.

[70] M. Li, X. Chen, X. Li, B. Ma, and M. B. Vitanyi, Paul, "The similarity metric," *IEEE Transactions on Information Theory*, vol. 50, no. 12, pp. 3250–3264, 2004.

[71] M. Beukman, S. James, and C. Cleghorn, "Towards objective metrics for procedurally generated video game levels," *arXiv preprint arXiv:2201.10334*, 2022.

[72] I. Levenshtein, Vladimir, "Binary codes capable of correcting deletions, insertions and reversals," *Soviet Physics Doklady*, vol. 10, p. 707, 1966.

[73] A. Zook, S. Lee-Urban, M. O. Riedl, H. K. Holden, R. A. Sottilare, and K. W. Brawner, "Automated scenario generation: Toward tailored and optimized military training in virtual environments," in *Proceedings of the International Conference on the Foundations of Digital Games*. ACM, 2012, pp. 164–171.

[74] M. Awiszus, F. Schubert, and B. Rosenhahn, "World-GAN: A generative model for Minecraft worlds," in *2021 IEEE Conference on Games*. IEEE, 2021, pp. 1–8.

[75] V. M. Chhieng and R. K. Wong, "Adaptive distance measurement for time series databases," in *Advances in Databases: Concepts, Systems and Applications*. Springer, 2007, pp. 598–610.

[76] Z. Wang, J. Liu, and G. N. Yannakakis, "The fun facets of Mario: Multifaceted experience-driven PCG via reinforcement learning," in *Proceedings of the 17th International Conference on the Foundations of Digital Games*. ACM, 2022, pp. 1–8.

[77] A. Alvarez, S. Dahlskog, J. Font, J. Holmberg, and S. Johansson, "Assessing aesthetic criteria in the evolutionary Dungeon designer," in *Proceedings of the 13th International Conference on the Foundations of Digital Games*. ACM, 2018, pp. 1–4.

[78] A. Alvarez, S. Dahlskog, J. Font, and J. Togelius, "Empowering quality diversity in Dungeon design with interactive constrained MAP-Elites," in *2019 IEEE Conference on Games*, 2019, pp. 1–8.

[79] ——, "Interactive constrained MAP-Elites: Analysis and evaluation of the expressiveness of the feature dimensions," *IEEE Transactions on Games*, vol. 14, no. 2, pp. 202–211, 2022.

[80] S. M. Lucas and V. Volz, "Tile pattern KL-divergence for analysing and evolving game levels," in *Proceedings of the Genetic and Evolutionary Computation Conference*, 2019, pp. 170–178.

[81] F. Schubert, M. Awiszus, and B. Rosenhahn, "TOAD-GAN: A flexible framework for few-shot level generation in token-based games," *IEEE Transactions on Games*, vol. 14, no. 2, pp. 284–293, 2022.

[82] X. Peng, J. C. Balloch, and M. O. Riedl, "Detecting and adapting to novelty in games," *arXiv preprint arXiv:2106.02204*, 2021.

[83] V. Bonometti, C. Ringer, M. Ruiz, A. Wade, and A. Drachen, "From theory to behaviour: Towards a general model of engagement," *arXiv preprint arXiv:2004.12644*, 2020.

[84] D. Melhart, A. Liapis, and G. N. Yannakakis, "The arousal video game annotation (AGAIN) dataset," *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 2171–2184, 2022.

[85] I. Karth and A. M. Smith, "Wave function collapse is constraint solving in the wild," in *Proceedings of the 12th International Conference on the Foundations of Digital Games*. ACM, 2017, pp. 1–10.

[86] R. R. Torrado, A. Khalifa, M. Cerny Green, N. Justesen, S. Risi, and J. Togelius, "Bootstrapping conditional GANs for video game level generation," in *2020 IEEE Conference on Games*. IEEE, 2020, pp. 41–48.

[87] M. Siper, S. Earle, Z. Jiang, A. Khalifa, and J. Togelius, "Controllable path of destruction," in *2023 IEEE Conference on Games*. IEEE, 2023, pp. 1–8.

[88] M. Siper, A. Khalifa, and J. Togelius, "Path of destruction: Learning an iterative level generator using a small dataset," in *2022 IEEE Symposium Series on Computational Intelligence*. IEEE, 2022, pp. 337–343.

[89] A. Khalifa, D. Perez-Liebana, S. M. Lucas, and J. Togelius, "General video game level generation," in *Proceedings of the Genetic and Evolutionary Computation Conference*. ACM, 2016, pp. 253–259.

[90] M. Frade, F. F. de Vega, and C. Cotta, "Evolution of artificial terrains for video games based on obstacles edge length," in *IEEE Congress on Evolutionary Computation*, 2010, pp. 1–8.

[91] ——, "Evolution of artificial terrains for video games based on accessibility," in *Applications of Evolutionary Computation*. Springer, 2010, pp. 90–99.

[92] ——, "Automatic evolution of programs for procedural generation of terrains for video games," *Soft Computing*, vol. 16, no. 11, pp. 1893–1914, 2012.

[93] N. Szilas and I. Ilea, "Objective metrics for interactive narrative," in *Interactive Storytelling*. Springer, 2014, vol. 8832, pp. 91–102.

[94] N. Szilas, "A computational model of an intelligent narrator for interactive narratives," *Applied Artificial Intelligence*, vol. 21, no. 8, pp. 753–801, 2007.

[95] N. Habonneau, U. Richle, N. Szilas, and J. E. Dumas, "3D simulated interactive drama for teenagers coping with a traumatic brain injury in a parent," in *Interactive Storytelling*. Springer, 2012, pp. 174–182.

[96] N. Partlan, E. Carstensdottir, S. Snodgrass, E. Kleinman, G. Smith, C. Harteveld, and M. S. El-Nasr, "Exploratory automated analysis of structural features of interactive narrative," in *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 14, no. 1. AAAI, 2018, pp. 88–94.

[97] A. Madkour, S. Marsella, and C. Harteveld, "Towards non-technical designer control over PCG systems: Investigating an example-based mechanism for controlling graph grammars," in *Proceedings of the 17th International Conference on the Foundations of Digital Games*. ACM, 2022, pp. 1–12.

[98] G. Smith, J. Whitehead, and M. Mateas, "Tanagra: A mixed-initiative level design tool," in *Proceedings of the Fifth International Conference on the Foundations of Digital Games*. ACM, 2010, pp. 209–216.

[99] S. Earle, M. Edwards, A. Khalifa, P. Bontrager, and J. Togelius, "Learning controllable content generators," in *2021 IEEE Conference on Games*, 2021, pp. 1–9.

[100] J. Mariño, W. Reis, and L. Lelis, "An empirical evaluation of evaluation metrics of procedurally generated Mario levels," in *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 11, no. 1. AAAI, 2015, pp. 44–50.

[101] A. Khalifa, M. C. Green, D. Perez-Liebana, and J. Togelius, "General video game rule generation," in *2017 IEEE Conference on Computational Intelligence and Games*, 2017, pp. 170–177.

[102] R. Marczak, G. Schott, and P. Hanna, "Understanding player experience through the use of similarity matrix," in *Proceedings of the 2015 DiGRA International Conference*, 2015, pp. 1–16.

[103] A. Coman and H. Muñoz-Avila, "Case-based plan diversity," in *Case-Based Reasoning. Research and Development*. Springer, 2010, pp. 66–80.

[104] A. Coman and H. Munoz-Avila, "Generating diverse plans using quantitative and qualitative plan distance metrics," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 25, no. 1. AAAI, 2011, pp. 946–951.

[105] D. Gravina, A. Liapis, and G. Yannakakis, "Surprise search: Beyond objectives and novelty," in *Proceedings of the Genetic and Evolutionary Computation Conference 2016*. ACM, 2016, pp. 677–684.

[106] A. Liapis, G. N. Yannakakis, and J. Togelius, "Enhancements to constrained novelty search: Two-population novelty search for generating game content," in *Proceedings of the 15th Annual Conference on Genetic and Evolutionary Computation*. ACM, 2013, pp. 343–350.

[107] D. Gravina, A. Liapis, and G. N. Yannakakis, "Quality diversity through surprise," *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 4, pp. 603–616, 2019.

[108] S. Sudhakaran, M. González-Duque, M. Freiberger, C. Glanois, E. Najarro, and S. Risi, "MarioGPT: Open-ended Text2Level generation through large language models," in *Thirty-Seventh Conference on Neural Information Processing Systems*, vol. 36, 2023, pp. 1–13.

[109] T. Shu, J. Liu, and G. N. Yannakakis, "Experience-driven PCG via reinforcement learning: A Super Mario Bros study," in *2021 IEEE Conference on Games*, 2021, pp. 1–9.

[110] M. Barthet, A. Liapis, and G. N. Yannakakis, "Open-ended evolution for Minecraft building generation," *IEEE Transactions on Games*, vol. 15, no. 4, pp. 603–612, 2023.

[111] A. S. Melotti and C. H. V. de Moraes, "Evolving roguelike Dungeons with deluged novelty search local competition," *IEEE Transactions on Games*, vol. 11, no. 2, pp. 173–182, 2019.

[112] G. Todd, S. Earle, M. U. Nasir, M. C. Green, and J. Togelius, "Level generation through large language models," in *Proceedings of the 18th International Conference on the Foundations of Digital Games*. ACM, 2023, pp. 1–8.

[113] E. C. Jackson and M. Daley, "Novelty search for deep reinforcement learning policy network weights by action sequence edit metric distance," in *Proceedings of the Genetic and Evolutionary Computation Conference Companion*. ACM, 2019, pp. 173–174.

[114] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.

[115] S. Snodgrass and S. Ontanon, "Procedural level generation using multi-layer level representations with MdMCs," in *2017 IEEE Conference on Computational Intelligence and Games*. IEEE, 2017, pp. 280–287.

[116] Z. Wang, J. Liu, and G. N. Yannakakis, "Keiki: Towards realistic Danmaku generation via sequential GANs," in *2021 IEEE Conference on Games*. IEEE, 2021, pp. 1–4.

[117] P. D. Sorensen, J. M. Olsen, and S. Risi, "Breeding a diversity of Super Mario behaviors through interactive evolution," in *2016 IEEE Conference on Computational Intelligence and Games*. IEEE, 2016, pp. 1–7.

[118] W. Sombat, P. Rohlfshagen, and S. M. Lucas, "Evaluating the enjoyability of the ghosts in Ms Pac-Man," in *2012 IEEE Conference on Computational Intelligence and Games*. IEEE, 2012, pp. 379–387.

[119] G. N. Yannakakis and J. Hallam, "Towards optimizing entertainment in computer games," *Applied Artificial Intelligence*, vol. 21, no. 10, pp. 933–971, 2007.

[120] ——, "Capturing player enjoyment in computer games," in *Advanced Intelligent Paradigms in Computer Games*. Springer, 2007, pp. 175–201.

[121] M. Cook, S. Colton, J. Gow, and G. Smith, "General analytical techniques for parameter-based procedural content generators," in *2019 IEEE Conference on Games*. IEEE, 2019, pp. 1–8.

[122] K. Zhang, J. Bai, and J. Liu, "Generating game levels of diverse behaviour engagement," in *2022 IEEE Conference on Games*. IEEE, 2022, pp. 167–174.

[123] K. Sorochan, J. Chen, Y. Yu, and M. Guzdial, "Generating Lode Runner levels by learning player paths with LSTMs," in *The 16th International Conference on the Foundations of Digital Games*, 2021, pp. 1–7.

[124] G. N. Yannakakis and J. Hallam, "A generic approach for obtaining higher entertainment in predator/prey computer games," *Journal of Game Development*, vol. 1, no. 3, pp. 23–50, 2005.

[125] J. Togelius, M. Preuss, and G. N. Yannakakis, "Towards multiobjective procedural map generation," in *Workshop on Procedural Content Generation in Games*. ACM, 2010, pp. 1–8.

[126] L. Ma, S. Cheng, M. Shi, and Y. Guo, "Angle-based multi-objective evolutionary algorithm based on pruning-power indicator for game map generation," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 6, no. 2, pp. 341–354, 2022.

[127] G. N. Yannakakis, J. Hallam, and H. H. Lund, "Comparative fun analysis in the innovative playware game platform," in *Proceedings of the 1st World Conference for Fun'n Games*. ACM, 2006, pp. 64–70.

[128] ——, "Capturing entertainment through heart rate dynamics in the playware playground," in *International Conference on Entertainment Computing*. Springer, 2006, pp. 314–317.

[129] ——, "Entertainment capture through heart rate activity in physical interactive playgrounds," *User Modeling and User-Adapted Interaction*, vol. 18, no. 1-2, pp. 207–243, 2008.

[130] C. Pedersen, J. Togelius, and G. N. Yannakakis, "Modeling player experience for content creation," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 2, no. 1, pp. 54–67, 2010.

[131] N. Shaker, J. Togelius, G. N. Yannakakis, B. Weber, T. Shimizu, T. Hashiyama, N. Sorenson, P. Pasquier, P. Mawhorter, G. Takahashi, G. Smith, and R. Baumgarten, "The 2010 Mario AI championship: Level generation track," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 3, no. 4, pp. 332–347, 2011.

[132] D. Loiacono, L. Cardamone, and P. L. Lanzi, "Automatic track generation for high-end racing games using evolutionary computation," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 3, no. 3, pp. 245–259, 2011.

[133] P. V. M. Dutra, S. M. Villela, and R. F. Neto, "Procedural content generation using reinforcement learning and entropy measure as feedback," in *21st Brazilian Symposium on Computer Games and Digital Entertainment*. IEEE, 2022, pp. 1–6.

[134] H. Simpson, Edward, "Measurement of diversity," *Nature*, vol. 163, no. 4148, pp. 688–688, 1949.

[135] M. Berland and V. Kumar, "Joint choice time: A metric for better understanding collaboration in interactive museum exhibits," in *13th International Learning Analytics and Knowledge Conference*. ACM, 2023, pp. 626–629.

[136] G. Smith, M. Cha, and J. Whitehead, "A framework for analysis of 2D platformer levels," in *Proceedings of the 2008 ACM SIGGRAPH Symposium on Video Games*. ACM, 2008, pp. 75–80.

[137] O. Withington and L. Tokarchuk, "The right variety: Improving expressive range analysis with metric selection methods," in *Proceedings of the 18th International Conference on the Foundations of Digital Games*. ACM, 2023, pp. 1–11.

[138] G. Smith, J. Whitehead, and M. Mateas, "Tanagra: An intelligent level design assistant for 2D platformers," in *Proceedings of the Sixth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*. AAAI, 2010, pp. 223–224.

[139] G. M. Smith, "Expressive design tools: Procedural content generation for game designers," Ph.D. dissertation, University of California, Santa Cruz, 2012.

[140] G. Smith, J. Whitehead, M. Mateas, M. Treanor, J. March, and M. Cha, "Launchpad: A rhythm-based level generator for 2-D platformers," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 3, no. 1, pp. 1–16, 2011.

[141] B. Horn, S. Dahlskog, N. Shaker, G. Smith, and J. Togelius, "A comparative evaluation of procedural level generators in the Mario AI framework," in *International Conference on the Foundations of Digital Games*. ACM, 2014, pp. 1–8.

[142] Y. Zakaria, M. Fayek, and M. Hadhoud, "Start small: Training controllable game level generators without training data by learning at multiple sizes," *Alexandria Engineering Journal*, vol. 72, pp. 479–494, 2023.

[143] A. Hald, J. S. Hansen, J. Kristensen, and P. Burelli, "Procedural content generation of puzzle games using conditional generative adversarial networks," in *International Conference on the Foundations of Digital Games*. ACM, 2020, pp. 1–9.

[144] N. Shaker, M. Nicolau, G. N. Yannakakis, J. Togelius, and M. O'Neill, "Evolving levels for Super Mario Bros using grammatical evolution," in *2012 IEEE Conference on Computational Intelligence and Games*. IEEE, 2012, pp. 304–311.

[145] S. Snodgrass and S. Ontanon, "A hierarchical MdMC approach to 2D video game map generation," in *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 11, no. 1, 2015, pp. 205–211.

[146] S. Snodgrass and S. Ontañón, "Controllable procedural content generation via constrained multi-dimensional Markov chain sampling," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. AAAI, 2016, pp. 780–786.

[147] ——, "Learning to generate video game maps using Markov models," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 9, no. 4, pp. 410–422, 2017.

[148] R. Van der Linden, R. Lopes, and R. Bidarra, "Designing procedurally generated levels," in *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 9, no. 3, 2013, pp. 41–47.

[149] W. Baghdadi, F. S. Eddin, R. Al-Omari, Z. Alhalawani, M. Shaker, and N. Shaker, "A procedural method for automatic generation of Spelunky levels," in *Applications of Evolutionary Computation*. Springer, 2015, pp. 305–317.

[150] J.-B. Hervé and C. Salge, "Comparing PCG metrics with human evaluation in Minecraft settlement generation," in *The 16th International Conference on the Foundations of Digital Games*. ACM, 2021, pp. 1–15.

[151] S. Dahlskog, J. Togelius, and M. J. Nelson, "Linear levels through N-Grams," in *Proceedings of the 18th International Academic MindTrek Conference: Media Business, Management, Content & Services*. ACM, 2014, pp. 200–206.

[152] S. Dahlskog and J. Togelius, "Procedural content generation using patterns as objectives," in *Applications of Evolutionary Computation*. Springer, 2014, pp. 325–336.

[153] S. Snodgrass, A. Summerville, and S. Ontanon, "Studying the effects of training data on machine learning-based procedural content generation," in *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 13, no. 1, 2021, pp. 122–128.

[154] D. Foreman-Mackey, "Corner.py: Scatterplot matrices in python," *Journal of Open Source Software*, vol. 1, no. 2, p. 24, 2016.

[155] M. Cook, J. Gow, and S. Colton, "Towards the automatic optimisation of procedural content generators," in *2016 IEEE Conference on Computational Intelligence and Games*, 2016, pp. 1–8.

[156] M. Cook, J. Gow, G. Smith, and S. Colton, "Danesh: Interactive tools for understanding procedural content generators," *IEEE Transactions on Games*, vol. 14, no. 3, pp. 329–338, 2022.

[157] M. Cook, J. Gow, and S. Colton, "Danesh: Helping bridge the gap between procedural generators and their output," in *Workshop on Procedural Content Generation at FDG*, 2016, pp. 1–16.

[158] J. Togelius, G. N. Yannakakis, K. O. Stanley, and C. Browne, "Search-based procedural content generation: A taxonomy and survey," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 3, no. 3, pp. 172–186, 2011.

[159] J. K. Pugh, L. B. Soros, and K. O. Stanley, "Quality diversity: A new frontier for evolutionary computation," *Frontiers in Robotics and AI*, vol. 3, no. 50, pp. 1–15, 2016.

[160] A. Khalifa, M. C. Green, G. Barros, and J. Togelius, "Intentional computational level design," in *Proceedings of the Genetic and Evolutionary Computation Conference*. ACM, 2019, pp. 796–803.

[161] L. B. Soros, J. K. Pugh, and K. O. Stanley, "Voxelbuild: A Minecraft-inspired domain for experiments in evolutionary creativity," in *Proceedings of the Genetic and Evolutionary Computation Conference Companion*. ACM, 2017, pp. 95–96.

[162] A. Khalifa, S. Lee, A. Nealen, and J. Togelius, "Talakat: Bullet hell generation through constrained MAP-Elites," in *Proceedings of the Genetic and Evolutionary Computation Conference*. ACM, 2018, pp. 1047–1054.

[163] M. Charity, M. C. Green, A. Khalifa, and J. Togelius, "Mech-Elites: Illuminating the mechanic space of GVG-AI," in *Proceedings of the 15th International Conference on the Foundations of Digital Games*. ACM, 2020, pp. 1–10.

[164] J. Schrum, B. Capps, K. Steckel, V. Volz, and S. Risi, "Hybrid encoding for generating large scale game level patterns with local variations," *IEEE Transactions on Games*, vol. 15, no. 1, pp. 46–55, 2023.

[165] M. C. Fontaine, R. Liu, A. Khalifa, J. Modi, J. Togelius, A. K. Hoover, and S. Nikolaidis, "Illuminating Mario scenes in the latent space of a generative adversarial network," in *AAAI Conference on Artificial Intelligence*, vol. 35. AAAI, 2021, pp. 5922–5930.

[166] K. Steckel and J. Schrum, "Illuminating the space of beatable Lode Runner levels produced by various generative adversarial networks," in *Proceedings of the Genetic and Evolutionary Computation Conference Companion*. ACM, 2021, pp. 111–112.

[167] V. R. Warriar, C. Ugarte, J. R. Woodward, and L. Tokarchuk, "Playmapper: Illuminating design spaces of platform games," in *2019 IEEE Conference on Games*. IEEE, 2019, pp. 1–4.

[168] M. Charity, A. Khalifa, and J. Togelius, "Baba is y'all: Collaborative mixed-initiative level design," in *2020 IEEE Conference on Games*. IEEE, 2020, pp. 542–549.

[169] M. C. Fontaine, S. Lee, L. B. Soros, F. De Mesentier Silva, J. Togelius, and A. K. Hoover, "Mapping Hearthstone deck spaces through MAP-Elites with sliding boundaries," in *Proceedings of the Genetic and Evolutionary Computation Conference*. ACM, 2019, pp. 161–169.

[170] M. C. Fontaine, J. Togelius, S. Nikolaidis, and A. K. Hoover, "Covariance matrix adaptation for the rapid illumination of behavior space," in *Proceedings of the 2020 Genetic and Evolutionary Computation Conference*. ACM, 2020, pp. 94–102.

[171] Y. Zhang, M. C. Fontaine, A. K. Hoover, and S. Nikolaidis, "Deep surrogate assisted MAP-Elites for automated Hearthstone deckbuilding," in *Proceedings of the Genetic and Evolutionary Computation Conference*. ACM, 2022, pp. 158–167.

[172] K. Sfikas, A. Liapis, and G. N. Yannakakis, "Monte Carlo elites: Quality-diversity selection as a multi-armed bandit problem," in *Proceedings of the Genetic and Evolutionary Computation Conference*. ACM, 2021, pp. 180–188.

[173] J. Togelius, M. Preuss, N. Beume, S. Wessing, J. Hagelbäck, and G. N. Yannakakis, "Multiobjective exploration of the StarCraft map space," in *Proceedings of the 2010 IEEE Conference on Computational Intelligence and Games*, 2010, pp. 265–272.

[174] J. Togelius, M. Preuss, N. Beume, S. Wessing, J. Hagelbäck, G. N. Yannakakis, and C. Grappiolo, "Controllable procedural map generation via multiobjective evolution," *Genetic Programming and Evolvable Machines*, vol. 14, no. 2, pp. 245–277, 2013.

[175] A. Khalifa and J. Togelius, "Multi-objective level generator generation with Marahel," in *Proceedings of the 15th International Conference on the Foundations of Digital Games*. ACM, 2020, pp. 1–8.

[176] H. T. T. Saurik, H. A. Rosyid, A. P. Wibawa, and E. I. Setiawan, "Evaluating player experience for fear modeling of 2D east Java horror game Alas Tilas," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 7, no. 4, pp. 858–864, 2023.

[177] I. S. Widiati, W. Hadi, M. Setiyawan, and Widada, "User experience evaluation of egrang traditional game application," in *2020 2nd International Conference on Cybernetics and Intelligent System*, 2020, pp. 1–5.

[178] P. Hämäläinen, J. Marshall, R. Kajastila, R. Byrne, and F. F. Mueller, "Utilizing gravity in movement-based games and play," in *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*. ACM, 2015, pp. 67–77.

[179] N. Partlan, E. Carstensdottir, E. Kleinman, S. Snodgrass, C. Harteveld, G. Smith, C. Matuk, S. C. Sutherland, and M. S. El-Nasr, "Evaluation of an automatically-constructed graph-based representation for interactive narrative," in *Proceedings of the 14th International Conference on the Foundations of Digital Games*. ACM, 2019, pp. 1–9.

[180] C. Holmgård, A. Liapis, J. Togelius, and G. N. Yannakakis, "Evolving models of player decision making: Personas versus clones," *Entertainment Computing*, vol. 16, pp. 95–104, 2016.

[181] K. Wejchert, *Elementy kompozycji urbanistycznej*. Wydawnictwo Arkady, 1984.

[182] J. Andrzejczak, M. Osowicz, and R. Szrajber, "Impression curve as a new tool in the study of visual diversity of computer game levels for individual phases of the design process," in *Computational Science – ICCS 2020*. Springer, 2020, vol. 12141, pp. 524–537.

[183] J. Schrum, J. Gutierrez, V. Volz, J. Liu, S. Lucas, and S. Risi, "Interactive evolution and exploration within latent level-design space of generative adversarial networks," in *Proceedings of the 2020 Genetic and Evolutionary Computation Conference*. ACM, 2020, pp. 148–156.

[184] N. Shaker, S. Asteriadis, G. N. Yannakakis, and K. Karpouzis, "Fusing visual and behavioral cues for modeling user experience in games," *IEEE Transactions on Cybernetics*, vol. 43, no. 6, pp. 1519–1531, 2013.

[185] A. Elor, M. Powell, E. Mahmoodi, M. Teodorescu, and S. Kurniawan, "Gaming beyond the novelty effect of immersive virtual reality for physical rehabilitation," *IEEE Transactions on Games*, vol. 14, no. 1, pp. 107–115, 2022.

[186] G. N. Yannakakis, H. H. Lund, and J. Hallam, "Modeling children's entertainment in the playware playground," in *2006 IEEE Symposium on Computational Intelligence and Games*, 2006, pp. 134–141.

[187] S. Yoo, C. Ackad, T. Heywood, and J. Kay, "Evaluating the actual and perceived exertion provided by virtual reality games," in *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 2017, pp. 3050–3057.

[188] S. Liu, C. Li, Y. Li, H. Ma, X. Hou, Y. Shen, L. Wang, Z. Chen, X. Guo, H. Lu, Y. Du, and Q. Tang, "Automatic generation of tower defense levels using PCG," in *Proceedings of the 14th International Conference on the Foundations of Digital Games*. ACM, 2019, pp. 1–9.

[189] V. Kumaran, B. Mott, and J. Lester, "Generating game levels for multiple distinct games with a common latent space," in *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 15, no. 1, 2019, pp. 102–108.

[190] W.-H. Huang, "Evaluating learners' motivational and cognitive processing in an online game-based learning environment," *Computers in Human Behavior*, vol. 27, no. 2, pp. 694–704, 2011.

[191] M. Nogueira, C. Cotta, and A. J. Fernández-Leiva, "On modeling, evaluating and increasing players' satisfaction quantitatively: Steps towards a taxonomy," in *Applications of Evolutionary Computation*. Springer, 2012, pp. 245–254.

[192] G. McAllister, P. Mirza-Babaei, and J. Avent, "Improving gameplay with game metrics and player metrics," in *Game Analytics: Maximizing the Value of Player Data*. Springer, 2013, pp. 621–638.

[193] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, J. Oh, D. Horgan, M. Kroiss, I. Danihelka, A. Huang, L. Sifre, T. Cai, J. P. Agapiou, M. Jaderberg, A. S. Vezhnevets, R. Leblond, T. Pohlen, V. Dalibard, D. Budden, Y. Sulsky, J. Molloy, T. L. Paine, C. Gulcehre, Z. Wang, T. Pfaff, Y. Wu, R. Ring, D. Yogatama, D. Wünsch, K. McKinney, O. Smith, T. Schaul, T. Lillicrap, K. Kavukcuoglu, D. Hassabis, C. Apps, and D. Silver, "Grandmaster level in StarCraft II using multi-agent reinforcement learning," *Nature*, vol. 575, no. 7782, pp. 350–354, 2019.

[194] A. Lupu, B. Cui, H. Hu, and J. Foerster, "Trajectory diversity for zero-shot coordination," in *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 2021, pp. 7204–7213.

[195] R. Shen, Y. Zheng, J. Hao, Z. Meng, Y. Chen, C. Fan, and Y. Liu, "Generating behavior-diverse game AIs with evolutionary multi-objective deep reinforcement learning," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. IJCAI, 2020, pp. 3371–3377.

[196] M. Szubert, W. Jaśkowski, P. Liskowski, and K. Krawiec, "The role of behavioral diversity and difficulty of opponents in coevolving game-playing agents," in *Applications of Evolutionary Computation*. Springer, 2015, pp. 394–405.

[197] R. Canaan, J. Togelius, A. Nealen, and S. Menzel, "Diverse agents for ad-hoc cooperation in Hanabi," in *2019 IEEE Conference on Games*. IEEE, 2019, pp. 1–8.

[198] E. Halina and M. Guzdial, "Diversity-based deep reinforcement learning towards multidimensional difficulty for fighting game AI," *arXiv preprint arXiv:2211.02759*, 2022.

[199] R. D. Gaina, R. Volkovas, C. G. Díaz, and R. Davidson, "Automatic game tuning for strategic diversity," in *2017 9th Computer Science and Electronic Engineering*, 2017, pp. 195–200.

[200] C. Hu, Y. Zhao, Z. Wang, H. Du, and J. Liu, "Game-based platforms for artificial intelligence research," *arXiv preprintarXiv:2304.13269*, 2023.

[201] L. Cardamone, G. N. Yannakakis, J. Togelius, and P. L. Lanzi, "Evolving interesting maps for a first person shooter," in *Applications of Evolutionary Computation*. Springer, 2011, vol. 6624, pp. 63–72.

[202] W. Cachia, A. Liapis, and G. Yannakakis, "Multi-level evolution of shooter levels," vol. 11, no. 1, 2021, pp. 115–121.

[203] K.-T. Chen and L.-W. Hong, "User identification based on game-play activity patterns," in *Proceedings of the 6th ACM SIGCOMM Workshop on Network and System Support for Games*. ACM, 2007, pp. 7–12.

[204] A. Saas, A. Guitart, and Á. Periáñez, "Discovering playing patterns: Time series clustering of free-to-play game data," in *2016 IEEE Conference on Computational Intelligence and Games*, 2016, pp. 1–8.

[205] A. T. Tekin, "Game level design in mobile gaming industry: Fuzzy pythagorean similarity approach," in *Intelligent and Fuzzy Systems*. Springer, 2022, pp. 676–683.

[206] R. Gallotta, G. Todd, M. Zammit, S. Earle, A. Liapis, J. Togelius, and G. N. Yannakakis, "Large language models and games: A survey and roadmap," *arXiv preprint arXiv:2402.18659*, 2024.

[207] J. Moizer, J. Lean, E. Dell'Aquila, P. Walsh, A. A. Keary, D. O'Byrne, A. Di Ferdinando, O. Miglino, R. Friedrich, R. Asperges, and L. S. Sica, "An approach to evaluating the user experience of serious games," *Computers & Education*, vol. 136, pp. 141–151, 2019.

[208] K. H. Koh, A. Basawapatna, V. Bennett, and A. Repenning, "Towards the automatic recognition of computational thinking for adaptive visual language learning," in *2010 IEEE Symposium on Visual Languages and Human-Centric Computing*, 2010, pp. 59–66.

[209] V. E. Bennett, K. Koh, and A. Repenning, "Computing creativity: Divergence in computational thinking," in *Proceeding of the 44th ACM Technical Symposium on Computer Science Education*. ACM, 2013, pp. 359–364.

[210] A. Basawapatna and A. Repenning, "Visualizing student game design project similarities," in *Diagrammatic Representation and Inference*. Springer, 2010, pp. 285–287.

SUPPLEMENTARY MATERIAL: FORMULATION OF DIVERSITY METRICS AND METHODS IN OBJECTIVE EVALUATION

This supplementary material presents the formulations of diversity metrics and methods used in objective evaluation.

For a comprehensive understanding of the notations used throughout this survey, we provide a detailed description of each symbol and its application in the context of evaluating game scenario diversity. These notations form the backbone of the mathematical models and analyses presented in the following sections.

Notations:

- $\mathcal{S}$: A set of scenarios or game levels under evaluation.
- x, x': Individual game scenarios or components within the set.
- d, d': The length of x, x', if the scenarios are represented in vector or sequence.
- P, Q: Probability or frequency distributions of considered items in scenarios, such as tile-pattern distribution.
- P_i, Q_i: The probability or frequency of the ith item in P and Q, respectively.
- $\boldsymbol{v}(\cdot)$: The featured vector of its argument.
- $v_i(\cdot)$: The ith entry of $\boldsymbol{v}(\cdot)$.
- $\delta(\cdot,\cdot)$: A comparison-based metric capturing the extent of dissimilarity or distance between two scenarios.
- $\phi(\cdot)$: A non-comparison-based metric summarizing some characteristic of scenarios into a value.

A. Formulations of metrics for objective evaluation

This section presents the formulations of diversity metrics used in objective evaluation. For a comprehensive introduction, citations, and discussions on related works regarding these metrics, please refer to Section V-A of the main manuscript.

1) Formulations of comparison-based metrics:

a) Distance Measures:

- Cosine similarity/distance

$$\delta(x, x') = \frac{\boldsymbol{v}(x) \cdot \boldsymbol{v}(x')}{\|\boldsymbol{v}(x)\|\|\boldsymbol{v}(x')\|} = \frac{\sum_{i=1}^{d} v_i(x)v_i(x')}{\sqrt{\sum_{i=1}^{d} v_i^2(x)}\sqrt{\sum_{i=1}^{d} v_i^2(x')}}.$$

- Manhattan distance

$$\delta(x, x') = \sum_{i=1}^{d} |v_i(x) - v_i(x')|.$$

- Euclidean distance

$$\delta(x, x') = \sqrt{\sum_{i=1}^{d} |v_i(x) - v_i(x')|^2}.$$

- Hamming distance

$$\delta(x, x') = \sum_{i=1}^{d} \mathbb{1}[v_i(x) \neq v_i(x')].$$

- Compression distance

$$\delta(x, x') = \frac{C(x \oplus x') - \min\{C(x), C(x')\}}{\max\{C(x), C(x')\}},$$

where $\oplus$ represents string concatenation, $C(\cdot)$ represents the length of its input after compression.

- Edit distance

$$\delta(x, x') = \delta_{d,d'}, \text{ with}$$

$$\forall 1 \leq i \leq d, \quad \delta_{i,0} = \sum_{k=1}^{i} w_{\text{del}}\left(v_k(x)\right)$$

$$\forall 1 \leq j \leq d', \quad \delta_{0,j} = \sum_{k=1}^{j} w_{\text{ins}}\left(v_k(x')\right),$$

$$\forall 1 \leq i, j \leq d', \quad \delta_{i,j} = \begin{cases} \delta_{i-1,j-1}, & \text{if } v_i(x) = v_j(x') \\ \min \begin{cases} \delta_{i-1,j} + w_{\text{del}}\left(v_i(x)\right) \\ \delta_{i,j-1} + w_{\text{ins}}\left(v_j(x')\right), \\ \delta_{i-1,j-1} + w_{\text{sub}}\left(v_i(x), v_j(x')\right) \end{cases} & \text{otherwise} \end{cases},$$

where w_{del}, w_{ins} and w_{sub} are predefined weights for deletion, insertion and substitution, repspectively.

- Dynamic time warping

$$\delta(x, x') = \delta_{d,d'}, \text{ with}$$
$$\delta_{i,j} = \Delta\left(v_i(x), v_j(x')\right) + \min\{\delta_{i-1,j}, \delta_{i,j-1}, \delta_{i-1,j-1}\},$$

where $\Delta(\cdot, \cdot)$ indicates a distance measure of two sequence items.

b) Divergence measures:

- Similarity

$$\delta(x|\mathcal{S}) = \sum_{i=1}^{r} \mathbb{1}[\exists y \in \mathcal{S}, v_i^{\text{row}}(x) = v_i^{\text{row}}(y)] + \sum_{i=1}^{c} \mathbb{1}[\exists y \in \mathcal{S}, v_i^{\text{col}}(x) = v_i^{\text{col}}(y)],$$

where $v_i^{\text{row}}(x)$ represents the feature value of ith row of x and $v_i^{\text{col}}(x)$ represents the feature value of the ith column of x. Typically, the feature used in this metric is the count of some kind of elements.

- KL divergence

$$\delta_{KL}(P\|Q) = \sum_{i=1}^{C} P_i \log \frac{Q_i}{P_i},$$

where C is the total number of categories.

- JS divergence

$$\delta_{JS}(P\|Q) = \frac{1}{2}\left(\delta_{KL}\left(P\left\|\frac{P+Q}{2}\right.\right) + \delta_{KL}\left(Q\left\|\frac{P+Q}{2}\right.\right)\right).$$

- Jaccard distance

$$\delta(\mathcal{S}, \mathcal{S}') = 1 - \frac{|\mathcal{S} \cap \mathcal{S}'|}{|\mathcal{S} \cup \mathcal{S}'|}.$$

c) Measures based on self-comparison:

- Symmetry

$$\delta(x) = \sum_{k=1}^{r/2} \mathbb{1}[v_{r/2-k}^{\text{row}}(x) = v_{r/2+k}^{\text{row}}(x)] + \sum_{k=1}^{c/2} \mathbb{1}[v_{c/2-k}^{\text{col}}(x) = v_{r/2+k}^{\text{col}}(x)],$$

where $v_i^{\text{row}}(x)$ represents the feature value of ith row of x and $v_i^{\text{col}}(x)$ represents the feature value of the ith column of x, r and c represent the number of rows and the number of columns, respectively.

2) Formulations of non-comparison-based metrics:

- Linearity

$$\phi(x) = \frac{1}{n} \sum_{i=1}^{n} |h_i(x) - L_i(x)|,$$

where n is the number of units, $h_i(x)$ denotes the height of the ith unit in the level x, and $L_i(x)$ represents the height of the best fit line for x's unit heights derived through linear regression.

- Leniency

$$\phi(x) = \frac{1}{kn} \sum_{i=1}^{n} \sum_{j=1}^{k} \text{score}(\#\text{item}_{i,j}),$$

where n is the number of units, and k is the count of item types considered, typically associated with the game's difficulty aspects, $\#\text{item}_{i,j}$ is the count of jth item in the ith column, $\text{score}(\cdot)$ is a scoring function based on the count of items in a column.

- Density

$$\phi(x) = \frac{\#\text{item}}{\#\text{total\_items}},$$

where $\#\text{item}$ and $\#\text{total\_item}$ are the count of the considered items and the count of all types of items, respectively.

B. Formulations of objective evaluation approaches

This section presents the formulations of objective evaluation approaches. For a comprehensive introduction, citations, and discussions on related works regarding these methods, please refer to Section V-B of the main manuscript.

1) Single-indicator methods: We use X to denote a generative scenario distribution to be evaluated (a set of scenarios can be viewed as a uniform distribution over the set).

a) Comparison-based methods:

- Average distance

$$D(X) = \mathbb{E}_{x \sim X, x' \sim X} \left[\delta(x, x') \right],$$

- Average nearest neighbor distance

$$D(X|\Re) = \mathbb{E}_{x \sim X} \left[\min_{x' \in \Re} \delta(x, x') \right], \tag{1}$$

where $\Re$ denotes a *reference set*, consisting of representative content samples, which can be the cluster centers selected by some clustering algorithm.

- Relative diversity

$$D(x|\mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{x' \in \mathcal{S}} \delta(x, x'), \tag{2}$$

- Novelty score

$$D(x|\mathcal{S}) = \frac{1}{k} \sum_{i=1}^{k} \delta(x, N_j(\mathcal{S})), \tag{3}$$

where $N_j(\mathcal{S})$ is the jth nearest neighbor in $\mathcal{S}$ in terms of δ, k is a parameter.

- Computational surprise

$$D(x|\mathcal{S}) = \frac{1}{k} \sum_{j=0}^{k} \delta(x, M_j(\mathcal{S})), \tag{4}$$

where M is a model that predicts multiple possible behaviors based on the history, and $M_j(\mathcal{S})$ indicates the j-closest prediction to x, k is a parameter.

- Slacking A-clipped function

$$D(x_i|x_{i-1} : x_{i-n}) = \frac{1}{R} \sum_{k=1}^{n} \min \left\{ 1 - \frac{|g - \delta(x_i, x_{i-k})|}{g}, r_k \right\}, \text{ with}$$

$$r_k = 1 - \frac{k}{n+1},$$

$$R = \sum_{k=1}^{n} r_k, \tag{5}$$

where n and g are hyperparameters, representing the number of previous segments to be considered and the desired value of divergence, respectively. In this metric, segments are sequential while x_i indicates the ith segment in the sequence.

b) Non-comparison-based methods:

- Standard deviation

$$D(X) = \sqrt{\mathbb{E}\left[\left(\phi(x) - \mathbb{E}[\phi(x)] \right)^2 \right]}, \tag{6}$$

- Behavior diversity

$$D(X) = \left(\frac{\sigma_\phi}{\sigma_{\max}} \right)^p, \text{ with}$$

$$\sigma_{\max} = \frac{1}{2} \sqrt{\frac{N}{N-1}} (\phi_{\max}(X) - \phi_{\min}(X)), \tag{7}$$

where σ_ϕ is the standard deviation over the N games regarding feature ϕ, $\phi_{\max(X)}$ and $\phi_{\min(X)}$ are the maximum and minimum values of ϕ across X, and p is a weighting parameter.

- Spatial diversity splits a level or map into some cells, and calculates the entropy over those cells as

$$D(x) = -\frac{1}{\log C} \sum_{i=1}^{C} \frac{\phi(x_i)}{\phi(x)} \log \frac{\phi(x_i)}{\phi(x)}, \tag{8}$$

where C is the total number of cells, x_i is the ith cell. Note the input x is lowercase because spatial diversity measures the intra-diversity of a single scenario. Notably, $\frac{\phi(x_i)}{\phi(x)}$ should be a frequency, i.e., $\sum_{i=1}^{C} \phi(x_i) = \phi(x)$ must be hold.

- Simpson index

$$D(x) = 1 - \sum_{i=1}^{C} \left(\frac{\phi(x_i)}{\phi(x)} \right)^2, \tag{9}$$

where C is the number of categories and $\frac{\phi(x_i)}{\phi(x)}$ should be a frequency, similar to spatial diversity.

2) Multi-indicator methods: ERA and QD analysis often rely on a partition Π of the feature (behavior) space $\mathcal{F}$, as defined by metrics in Section V-A. A partition of $\mathcal{F}$ is a collection of subsets of $\mathcal{F}$ such that the union of all items in Π equals to $\mathcal{F}$, and any two items π and π' are non-overlapping, i.e., $\pi \cap \pi' = \varnothing$. In the context of multi-indicator approaches, the evaluation involves a set of samples $\mathcal{S}$ indexed by the partition Π. We denote the samples within $\mathcal{S}$ corresponding to a specific partition item π as $\mathcal{S}(\pi)$, expressed as $\mathcal{S}(\pi) = \{x | x \in \mathcal{S} \wedge \phi(x) \in \pi\}$. It's worth noting that the Π can be a non-uniform partition.

- QD score

$$\mathrm{QDS}(\mathcal{S}) = \frac{1}{|\Pi|} \sum_{\pi \in \Pi} \max_{x \in \mathcal{S}(\pi)} q(x), \tag{10}$$

  where $q(x)$ is the quality score of x.
- Coverage

$$\mathrm{Cov}(S) = \frac{1}{|\Pi|} \sum_{\pi \in \Pi} \mathbb{1}[\mathcal{S}(\pi) \neq \varnothing]. \tag{11}$$

- Hypervolume

$$\mathrm{HV}(X|\mathbf{r}) = \Lambda \left(\bigcup_{x \in X} \{p \in \mathbb{R}^m \mid p \succ \mathbf{r} \wedge \boldsymbol{v}(x) \succ p\} \right), \tag{12}$$

  where $\Lambda(\cdot)$ indicates the Lebesgue measure, m is the number of objectives, $\succ$ indicates dominating, $\boldsymbol{v}(x)$ is the vector of x's all objective values, and $\mathbf{r}$ is a m-dimensional reference point.